Monocular 3D Pose Estimation via Pose Grammar and Data Augmentation

Yuanlu Xu, Wenguan Wang, Tengyu Liu, Xiaobai Liu, Jianwen Xie and Song-Chun Zhu

Abstract—In this paper, we propose a pose grammar to tackle the problem of 3D human pose estimation from a monocular RGB image. Our model takes estimated 2D pose as the input and learns a generalized 2D-3D mapping function to leverage into 3D pose. The proposed model consists of a base network which efficiently captures pose-aligned features and a hierarchy of Bi-directional RNNs (BRNNs) on the top to explicitly incorporate a set of knowledge regarding human body configuration (*i.e.*, kinematics, symmetry, motor coordination). The proposed model thus enforces high-level constraints over human poses. In learning, we develop a data augmentation algorithm to further improve model robustness against appearance variations and cross-view generalization ability. We validate our method on public 3D human pose benchmarks and propose a new evaluation protocol working on cross-view setting to verify the generalization capability of different methods. We empirically observe that most state-of-the-art methods encounter difficulty under such setting while our method can well handle such challenges.

Index Terms—3D pose estimation, dependency grammar, data augmentation, deep neural network, recurrent neural network, evaluation protocol, learning-by-synthesis

1 Introduction

Estimating 3D human poses from a single-view RGB image has attracted growing interest in the past few years for its wide applications in robotics, autonomous vehicles, intelligent drones, *etc*. This is a challenging inverse task since it aims to reconstruct 3D spaces from 2D data and the inherent ambiguity is further amplified by other factors, *e.g.*, clothes, occlusions, background clutters. With the availability of large-scale pose datasets, *e.g.*, *Human3.6M* [1], deep learning based methods have obtained encouraging success. These methods can be roughly divided into two categories: i) learning end-to-end networks that recover 2D input images to 3D poses directly, ii) extracting 2D human poses from input images and then lifting 2D poses to 3D spaces.

There are some advantages to decouple 3D human pose estimation into two stages. i) For 2D pose estimation, existing large-scale pose estimation datasets [2], [3] have provided sufficient annotations; whereas pre-trained 2D pose estimators [4] are also generalized and mature enough to be deployed elsewhere. ii) For 2D to 3D reconstruction, infinite 2D-3D pose pairs can be generated by projecting each 3D pose into 2D poses under different camera views. Recent work [5], [6] have shown that well-designed deep networks can achieve state-of-the-art performance on *Human3.6M* dataset using only 2D pose detections as system inputs.

However, despite their promising results, few previous methods explored the problem of encoding domain-specific knowledge into current deep learning based detectors.

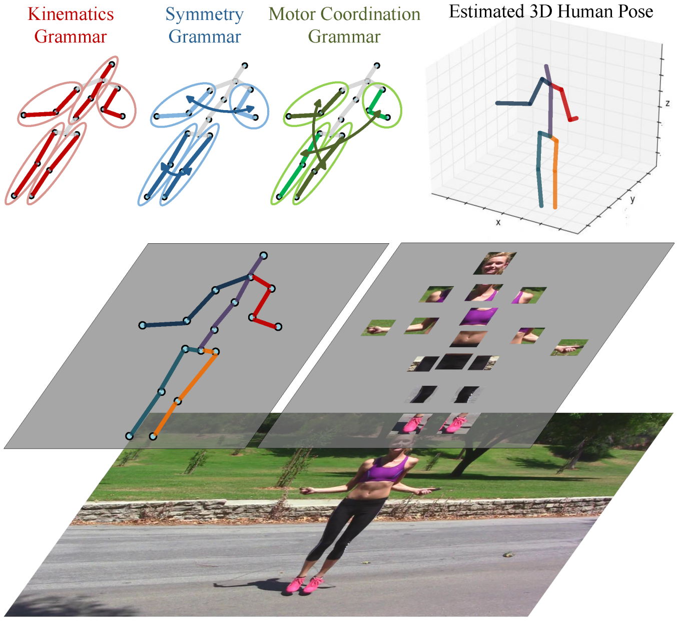


Fig. 1. Illustration of human pose grammar, which expresses the knowledge of human body configuration. We consider three kinds of human body dependencies and relations in this paper, *i.e.*, kinematics (red), symmetry (blue) and motor coordination (green).

In this paper, we develop a deep grammar network to encode low-level appearance and geometry features as well as high-level knowledge over human body dependencies and relations, as illustrated in Figure 1. Such knowledge explicitly expresses the composition process of joint-part-pose, including kinematics, symmetry and motor coordination, and serve as knowledge bases for reconstructing 3D poses. We ground such knowledge in a multi-level RNN network

- *Y. Xu, W. Wang, T. Liu and S.-C. Zhu are with the Department of Computer Science and Statistics, University of California, Los Angeles (UCLA), Los Angeles, CA 90095, USA. E-mail: merayxu@gmail.com, wenguanwang.ai@gmail.com, tengyuliu@cs.ucla.edu, sczhu@stat.ucla.edu*
- *X. Liu is with the Department of Computer Science, San Diego State university (SDSU), San Diego, CA 92182, USA. E-mail: xiaobai.liu@mail.sdsu.edu*
- *J. Xie is with Baidu Research, USA. E-mail: jianwen@ucla.edu*

which can be end-to-end trained with back-propagation. The composed hierarchical structure describes composition, context and high-order relations among human body parts.

Additionally, we empirically find that previous methods are restricted to their poor generalization capabilities while performing cross-view pose estimation, *i.e.*, being tested on human images from unseen camera views. Notably, on the *Human3.6M* dataset, the largest publicly available human pose benchmark, we find that the performance of state-of-the-art methods heavily relies on the camera viewpoints. As shown in Table 1, once we change the split of training and testing set, using three cameras for training and testing on the forth camera (new *Protocol #4*), performance of state-of-the-art methods drops dramatically and is much worse than image-based deep learning methods. These empirical studies suggested that existing methods might over-fit to sparse camera settings and bear poor generalization capabilities.

To handle the issue, we propose to augment the learning process in two aspects: i) enriching the appearance model with in-the-wild scenarios, and ii) incorporating 2D poses under more camera views. The proposed data augmentation algorithm explores a more robust and generalized mapping from 2D poses to 3D poses. More specifically, we develop a pose simulator to augment training samples with virtual camera views, which can further improve system robustness. Our method is motivated by the previous work on learning by synthesis. In contrast, we focus on sampling of 2D pose instances re-projected from a given 3D pose configuration, following the basic geometry principles. In particular, we develop a pose simulator to effectively generate training samples from unseen camera views. These samples can greatly reduce the risk of over-fitting and thus improve generalization capabilities of the developed pose estimation system.

We conduct quantitative experiments on three public 3D human pose benchmarks, namely *Human3.6M* [1], *HumanEva* [7], and *HHOI* [8] to evaluate the proposed method for cross-view human pose estimation. Additionally, qualitative experiments in *MPII* [2] dataset, using our model learned on *Human3.6M*, demonstrate compelling results and good generalization capability on in-the-wild images. Furthermore, to gain a complete and comprehensive understanding of its various aspects, we implement three variants of our method and conduct ablative studies. Results show that our method can significantly reduce pose estimation errors and outperform the alternative methods to a large extend.

Contributions. To summarize, our method has the following contributions:

- A deep grammar network that incorporates both powerful encoding capabilities of deep neural networks and high-level dependencies and relations of human body.
- A data augmentation algorithm that improves model robustness against appearance variations and cross-view generalization ability.
- A new evaluation protocol for validating model cross-view generalization abilities on *Human3.6M* dataset.

This paper builds upon two earlier conference paper [9], [10]. In [9], we proposed a pose grammar that mainly accounts kinematic structure encoded in human body and is learned by a tree-LSTM scheme. Then, in [10], we consolidated our pose estimator as an amalgamation of three dependency grammar: kinematics, symmetry, and motor coordination, and presented a BRNN-based learning framework for efficiently capturing the bi-directional relations in such dependency grammar. The present work significantly extends previous studies with integration of both appearance and geometry feature encoders, an improved data augmentation algorithm for the new model, and in-depth discussions and more enclosed details about the proposed framework. Additionally, it presents a more comprehensive analysis of the proposed testing protocol, which provides an essential way for evaluating the overfitting with camera. It also offers a more inclusive and insightful overview of the recent work of 3D human pose estimation. Last but not least, it reports extensive experimental results with deeper analysis.

The remainder of the paper is organized as follows. An overview of the related work is presented in Section 2. Section 3 explains dependency grammar based 3D human pose modeling. The learning process with the proposed pose sample simulator is described in Section 5. In Section 6, we offer detailed experimental analyses on robustness, effectiveness, and efficiency. Finally, we draw conclusions in Section 7.

2 Related Work

The proposed method is closely related to two streams of research in the literature, *i.e.*, monocular 3D pose estimation (Section 2.1), and grammar model (Section 2.2), which we will briefly in the following.

2.1 Monocular 3D Pose Estimation

Estimating 3D human poses [11], [12] has been extensively studied for the past few years, as surveyed in [13]. Among which, people tackle this tasks under different settings, such as multi-view estimation [14], [15], [16], [16], video based estimation [17], [18], [19] and single image based estimation [20], [21], [22].

Our method falls into the third setting. Unlike multi-view and video settings which could utilize large amount of complementary information from different camera views and motion in the temporal domain, our method, as well as other single image based estimators solely rely on leveraging 2D image features into the 3D space. Thanks to powerful deep neural networks developed recently, lots of deep learning based methods have been proposed recently. Methods in the literature can be generally classified into two categories: i) directly learning 3D pose structures from 2D images [23], [24], [25], [26], [27], ii) a cascaded framework of first performing 2D pose estimation and then reconstructing 3D pose from the estimated 2D joints.

Specifically, for the first class, Li *et al.* [28] regressed 3D pose directly from images using convolutional network. Then, in [29], they exploited an image-pose embedding network to regularize the 3D pose structures, which is trained using a maximum-margin cost function. Tekin *et al.* [30] first learned an auto-encoder that describes 3D pose in high dimensional space and then mapped the input image to that space using CNN. Pavlakos *et al.* [31] represented 3D joints as points in a discretized 3D space and proposed a coarse-to-fine approach for iterative refinement. More recently, Sun *et al.* [32] proposed a structure-aware network that regresses bones instead of joint as pose representations and exploits a compositional loss function that encodes interactions between the bones. Their training and testing are restricted to the 3D Motion Capture data in constrained environments, since it is difficult to obtain large-scale 3D data for arbitrary pose and the collection relies on professional tools for 3D marker tracking.

To alleviate above limitations, some work [5], [33], [6], [34], [35], [36], [37] tried to address this problem in a two step manner that reasons 3D pose through intermediate 2D joints predictions from other off-the-shelf methods. The advantage lies on modular training, where 2D datasets (typically larger and more diverse due to the ease of annotation) can be used to train the initial visual analysis module, while 3D motion capture data (difficult to obtain and limited to controlled lab environments) can be used to train the subsequent 3D reasoning module. More specially, in [5], the authors proposed an exemplar-based method to retrieve the nearest 3D candidate from a 3D pose library using the estimated 2D landmarks. Zhou *et al.* [17] predicted 3D poses from a video sequence by using temporal information. The estimation process is conducted via an EM type algorithm over the entire sequence and 2D joint uncertainties are also marginalized out during inference. In [33], they mixed 2D and 3D data and trained a two-stage cascaded network via considering extra 3D geometric constrains. Martinez *et al.* [6] proposed an efficient network architecture that directly regresses 3D keypoints from 2D joint detections and demonstrates most of the error of current deep 3D pose estimators stems from visual analysis module. More recently, people also tried to model the ambiguities of lifting 2D to 3D, using Multiple Hypothesis Learning [35] and Generative Adversarial Network [38]. In [36], authors proposed to disentangle feature encodings of explicit 2D and 3D features and thus enabled them to perform joint training on both supervised and unsupervised data.

In this paper, we propose a unified 2D-to-3D reconstruction network that integrates the learning power of deep learning and the domain-specific knowledge represented by hierarchical grammar model. Different from [37] which learns semantic relationships using Graph Convolution Network (GCN) in an unsupervised way, our grammar rules are designed based on heuristics. This significantly reduces model ambiguities and overfitting to dataset bias.

2.2 Grammar Model

Grammar models, originated in neural language [39], have received long-lasting endorsement [40], [41] in computer vision and robotics, due to its expressive power in modeling structures and relations, which are ubiquitous in above research areas. The grammar models can be categorized into two principal variations [39]: phrase structure grammar and dependency grammar.

The **phrase structure grammar** is based on the constituency relation. The constituency relation defines the rule to break down a node (*e.g.*, parent node) into its constituent parts (*e.g.*, child nodes). In other words, each node must geometrically contain all of its constituents [42], [43]. Phrase structure grammar were introduced in syntactic pattern recognition by K.S. Fu in the early 1980s [44], and rejuvenated into compositional models by Geman [45], and stochastic and-or grammar by Zhu and Mumford [46]. The advantage of the phrase structure grammar lies in its coarse-to-fine summarization ability, thus it has been widely used for modeling the compositional structures in object recognition [42], [47], [48], [49], [50], scene parsing [51], [52], [53] and event understanding [54], [55].

In this work we focus on **dependency grammar**, where constituent parts do not need to be contained within their parents, but instead are constrained by an adjacency relation [56]. Thus the dependency grammar is well suited for representing articulated relations among human parts [57]. In computer vision, the pictorial model [58] and the flexible mixture of parts model [59] can be viewed as dependency grammar. Our work extends dependency grammar with deep learning for monocular 3D pose estimation, where the dependency grammar acted as global constraints for more reasonable 3D layout estimation. More specially, in our early version [9], we represented human body as a set of simplified dependency grammar based on kinematic relations and learn the grammar with tree-LSTM. Then, in [10], we presented a more general and unified dependency grammar model that learns kinematic, symmetric, and coordinating relations within hierarchical human configurations using Bidirectional RNNs.

3 Method Overview

In this section, we firstly introduce the problem formulation of the task of monocular 3D pose estimation (Section 3.1) and then briefly overview the concept of pose grammar used in our framework (Section 3.2).

3.1 Problem Formulation

Given an image $\mathbf{I}$, with some off-the-shelf 2D pose estimators, we can obtain the detected 2D human pose $\mathbf{U}$, represented as a set of N_v joint locations,

$$\mathbf{U} = \{u_i : i = 1, \dots, N_v,\ u_i \in \mathbb{R}^2\}. \tag{1}$$

Our task is to estimate the corresponding 3D human pose $\mathbf{V}$ in the world reference frame,

$$\mathbf{V} = \{v_i : i = 1, \dots, N_v,\ v_i \in \mathbb{R}^3\}. \tag{2}$$

Suppose the coordinate of 2D joint $u_i = [x_i, y_i]$ and the coordinate of 3D joint $v_i = [X_i, Y_i, Z_i]$, we can describe the relation between 2D and 3D as a perspective projection, that is,

$$\begin{bmatrix} x_i \\ y_i \\ w_i \end{bmatrix} = K[R|RT] \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix}, K = \begin{bmatrix} \alpha_x & 0 & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{bmatrix}, T = \begin{bmatrix} T_x \\ T_y \\ T_z \end{bmatrix}, \tag{3}$$

where w_i is the depth w.r.t. the camera reference frame, K is the camera intrinsic parameter (*e.g.*, focal length α_x and α_y, principal point x_0 and y_0), R and T are camera extrinsic parameters of rotation and translation, respectively. Notably, we omit camera distortion to simplify the imaging principals.

There exist two sub-problems in estimating 3D pose from 2D pose: i) calibrating camera parameters, and ii) estimating 3D human joint positions. Noticing that these two sub-problems are entangled and cannot be solved without ambiguity, we propose a deep neural network to learn the generalized 2D to 3D mapping

$$\mathbf{V} = f(\mathbf{I}, \mathbf{U}; \theta), \tag{4}$$

where $f(\cdot)$ denotes a multi-to-multi mapping function from 2D image space to 3D space, taken both image appearance and 2D geometry as input and parameterized by θ.

In this paper, we regard the 2D pose estimation results as the weakly-supervised geometry information in localizing and aligning the articulated human body. The given 2D poses can be obtained from any off-the-shelf 2D pose estimation methods. The major benefit of such formulation lies in that we could plug-in and make full usage of advances from state-of-the-arts 2D pose estimators and there is no need to re-train the whole framework.

In this paper, we follow the pipeline proposed in [9] which takes images and 2D poses as joint inputs, instead of the framework proposed in [10] which isolates image input from 3D pose estimation. Specifically, we extract image patches $\mathbf{I}_U = \{\mathbf{I}_{u_i}, i = 1, \ldots, N_v\}$ given the 2D pose estimation results. $\mathbf{I}_U$ is regarded as additional inputs, which offers complementary visual information besides the detected 2D joints $\mathbf{U}$. To align 3D poses $\mathbf{V}$, we consider the *pelvis* joint as the origin in the 3D coordinate system and thus transform absolute 3D coordinates for each joints into relative 3D coordinates by subtracting the absolute 3D coordinates of the *pelvis* joint.

3.2 Pose Grammar

The human body nature which involves rich inherent structures motivates us to reason the 3D structure of the whole person in a high-level, structural manner. Before going deep into our grammar based 3D pose estimator, we first detail three essential kinds of human body structures, namely *kinematics*, *symmetry*, and *motor coordination* grammar, which reflect interpretable and high-level knowledge of human body configuration.

Kinematics grammar $\mathcal{G}^{kin}$ describes human body movements without considering forces. Such structures represent motion properties of human in a topological manner. Kinematics grammar focuses on connected body parts and works both forward and backward. Forward kinematics takes the last joint in a kinematic chain into account while backward kinematics reversely influences a joint in a kinematics chain from the next joint.

Symmetry grammar $\mathcal{G}^{sym}$ measures mirror symmetry of human body, as human body can be divided into matching halves by drawing a line down the center; the left and right sides are roughly mirror images of each other. Actually, such mirror symmetry, also named bilateral symmetry, is an important principle in nature; vast majority of animals (more than 99%) are bilaterally symmetric, including humans [60].

Motor coordination grammar $\mathcal{G}^{crd}$ represents movements of several limbs combined in a certain manner. More specially, motor coordination can be thought as each physiological process created with the kinematic and kinetic parameters that must be performed in order to achieve intended actions. It can be seen everywhere, involved in moving a limb to picking up a ball to shooting the ball.

4 Network Architecture

As illustrated in Figure 2, our model follows the line that estimating 3D human keypoints from both input RGB images and intermediate 2D joint detections, which renders our model highly applicable. More specifically, we extend several human pose grammar rules into the deep neural network, where a base 3D pose estimation network is first used for extracting pose-aligned features, and a hierarchy of RNNs is built for encoding high-level 3D pose grammar for generating final reasonable 3D pose estimations. Above two networks work in a cascaded way, resulting in a strong 3D pose estimator that inherits the representation power of neural network and high-level knowledge of human body configuration.

4.1 Base 3D Pose Network

To build a solid foundation for high-level grammar model, we first use a base network for capturing both 2D and 3D well pose-aligned features. In this paper, we mainly encode two types of features from the input: appearance features from input images and geometry features from 2D poses.

4.1.1 *Appearance Feature Encoding*

Given an input image $\mathbf{I}$ and 2D poses $\mathbf{U}$, we first sample 56×56 image patches $\mathbf{I}_U$ centered around each 2D joints and extract appearance features by feeding them into a convolutional neural network. Our appearance feature encoder is mainly inspired by ResNet [61]. The convolutional layers mostly have 3×3 filters and the network follows two rules in design: i) the layers share the same number of filters for the same output feature map size; (ii) the number of filters will double if the feature map shrinks by half, in order to preserve the per-layer time complexity. The feature map shrinking is obtained by having a stride of 2 in the convolutional layers.

As illustrated in Figure 2 bottom left, each image patch goes through 4 cascaded blocks. Each block consists of stacks of convolutional layers, Batch Normalization and ReLU activation. We keep the same size of the output feature map as the input for block 1 and downsample that to half in each of the last three blocks by setting the stride as 2 in the first convolutional layer. Therefore, feature maps obtained by each blocks are with size of $56 \times 56 \times 64$, $28 \times 28 \times 128$, $14 \times 14 \times 256$ and $7 \times 7 \times 512$, respectively. The network ends with a global average pooling layer and outputs a 512-*d* feature vector. Note layer weights are shared across all N_v joints in the encoder and all extracted features are concatenated together into a long vector.

4.1.2 *Geometry Feature Encoding*

The geometry feature encoder is inspired by [6], which has been demonstrated to be effective in encoding the information of 2D and 3D poses. As illustrated in Figure 2, our base network consists of two cascaded blocks. For each block, several linear (fully connected) layers, interleaved

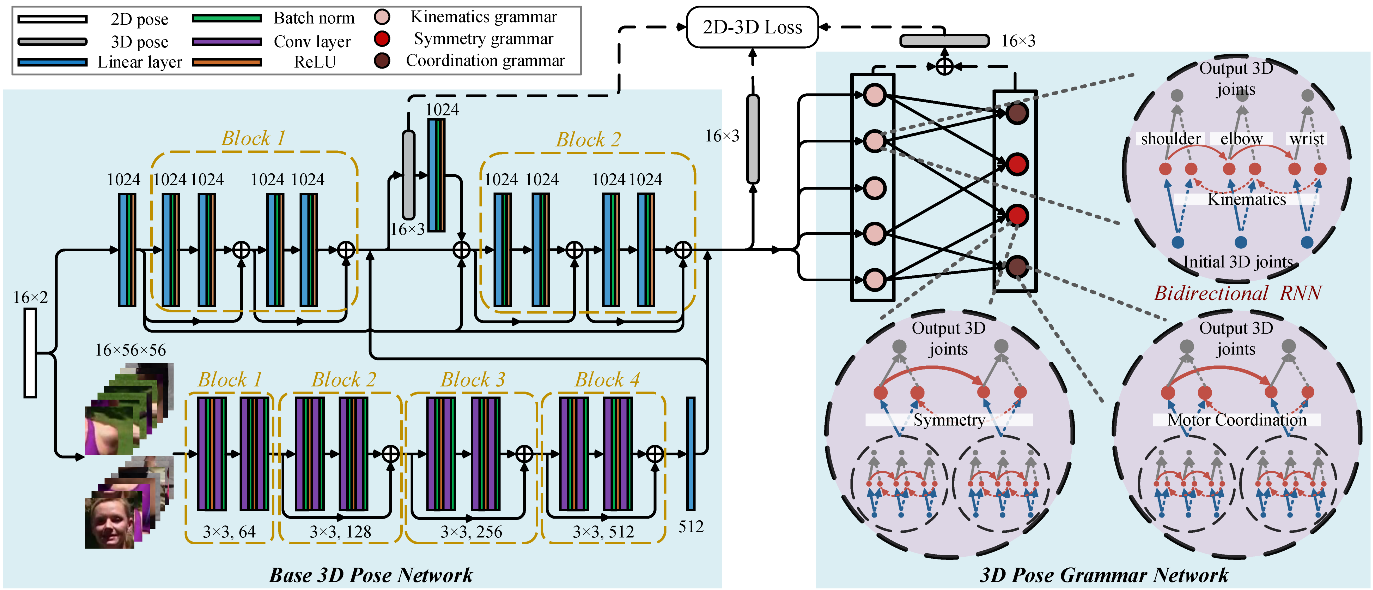


Fig. 2. The proposed deep grammar network. Our model consists of two major components: i) a base network constituted by two submodules encoding appearance and geometry features given the input image and the detected 2D pose, ii) a pose grammar network encoding human body dependencies and relations w.r.t. kinematics, symmetry and motor coordination. Each grammar is represented as a Bi-directional RNN among certain joints. The current state of a joint is composed of the hidden states of its related joints and the encoded input feature from itself. See text for detailed explanations.

with Batch Normalization, ReLU activation and Dropout layers, are stacked for efficiently mapping the 2D pose features to higher-dimensions.

As illustrated in Figure 2 top left, the input 2D pose detections $\mathbf{U}$ (obtained as ground-truth 2D joint locations under known camera parameters, or from other 2D pose detectors) are first projected into a 1024-*d* features, with a fully connected layer. Then the first block takes this high-dimensional features as input and an extra linear layer is applied at the end of it to obtain an explicit 3D pose representation. In order to have a coherent understanding of the full body in 3D space, we re-project the 3D estimation into a 1024-*d* space and further feed it into the second block. With the initial 3D pose estimation from the first block, the second block is able to reconstruct a more reasonable 3D pose.

To make full use of the information from initial 2D pose detections, we utilize *residual connections* between the two blocks. Such technique is able to encourage the information flow and facilitate our training. Additionally, each block in our base network is able to directly access to the gradients back-propagated from the loss function (detailed in Section 5), leading to an implicit deep supervision [62]. With the refined 3D pose estimated from base network, we again re-projected it into a 1024-*d* features. We combine the 1024-*d* features from the 3D pose and the original 1024-*d* feature of 2D pose together, which leads to a final 1024-*d* feature representation that has well-aligned 3D pose information and preserves the original 2D pose information.

Finally, we concatenate appearance and geometry feature representations together and feed this long vector into our 3D pose grammar network.

4.2 3D Pose Grammar Network

So far, our base network directly estimated the depth of each joint from the 2D pose detections. However, such end-to-end estimation ignores the rich inherent structures within the human body. Motivated by this, we propose to incorporate reasoning of the 3D human body structure in a global manner. Here we use Recurrent Neural Network (RNN) to model high-level knowledge of 3D human pose grammar, which towards a more reasonable and powerful 3D pose estimator that is capable of satisfying human anatomical and anthropomorphic constraints. Before going deep into our grammar network, we first detail our grammar formulations that reflect interpretable and high-level knowledge of human body configuration. Basically, given a human body, we consider the following three types of grammar in our network.

In this paper, we derive pose grammar on a human skeleton structure of 17 joints defined in *Human3.6M* [1], *i.e.*, head, neck, thorax, spine, pelvis, shoulders, elbows, wrists, hips, knees, and ankles. As we mentioned in Section 3.1, we shift all other joints to a relative coordinate system centered by the *pelvis* and thus only 16 joints are represented in the proposed pose grammar. The used skeleton structure can be easily adapted to other skeleton structures (*e.g.*, LSP format 14 joints [63], MPII format 15 joints [64], Kinect format 13 joints) by skipping unused joints.

Kinematics grammar $\mathcal{G}^{kin}$. Kinematic constraints are an essential factor for building a human body representation. There are numerous human body simulators [65], human dynamics tracking systems [66], [67] and tree-based pose estimation methods [68], [69] are built upon kinematic structures. In this paper, we represent the kinematics within the human body as a tree structure. The articulated relation are better represented and the correlation of features at parent joint and child joint are better captured within tree structure than the flat or sequential structure. Similar to the framework of [70], we adapt the tree-structured recurrent neural network for modeling human pose and integrating local and global features. The aggregated contextual information are propagated efficiently through the edges between joints.

To simplify the representation, we decompose the human body tree structure into multiple chains with overlapping nodes. As illustrated in Figure 1 the red skeleton, five chains (*i.e.*, kinematics grammar) are defined to represent the kinematic constraints within different body parts:

$$\begin{aligned}
\mathcal{G}^{kin}_{vert} &: \textit{head} \leftrightarrow \textit{neck} \leftrightarrow \textit{thorax} \leftrightarrow \textit{spine}\ , && (5)\\
\mathcal{G}^{kin}_{l.arm} &: \textit{thorax} \leftrightarrow \textit{l.shoulder} \leftrightarrow \textit{l.elbow} \leftrightarrow \textit{l.wrist}\ , && (6)\\
\mathcal{G}^{kin}_{r.arm} &: \textit{thorax} \leftrightarrow \textit{r.shoulder} \leftrightarrow \textit{r.elbow} \leftrightarrow \textit{r.wrist}\ , && (7)\\
\mathcal{G}^{kin}_{l.leg} &: \textit{spine} \leftrightarrow \textit{l.hip} \leftrightarrow \textit{l.knee} \leftrightarrow \textit{l.foot}\ , && (8)\\
\mathcal{G}^{kin}_{r.leg} &: \textit{spine} \leftrightarrow \textit{r.hip} \leftrightarrow \textit{r.knee} \leftrightarrow \textit{r.foot}\ . && (9)
\end{aligned}$$

Our kinematic tree model is powerful as it captures the most important source of constraint on human body pose. However, this model is also limited by the fact that it does not represent information about relations between limbs that are not physically connected by bones. Therefore, we further incorporate two advanced relations, *i.e.*, symmetry and motor coordination, into the pose grammar. We organize these two advanced grammar on top of the kinematics grammar; that is, symmetry and motor coordination grammar network takes outputs from the kinematics grammar network as inputs and further optimizes 3D pose configurations.

Symmetry grammar $\mathcal{G}^{sym}$. Human body has a symmetrical appearance when viewed externally, *i.e.*, the two halves of human body are nearly mirror images of each other, when a line is drawn in the center. For decades, people explore various computational models [71], [72], [73], analytical or biological basis, proving the effectiveness in computer vision and graphics.

As illustrated in Figure 1 the blue skeleton, we build two symmetry grammar for describing such bilateral symmetrical properties of human body:

$$\begin{aligned}
\mathcal{G}^{sym}_{arm} &: \mathcal{G}^{kin}_{l.arm} \leftrightarrow \mathcal{G}^{kin}_{r.arm}\ , && (10)\\
\mathcal{G}^{sym}_{leg} &: \mathcal{G}^{kin}_{l.leg} \leftrightarrow \mathcal{G}^{kin}_{r.leg}\ . && (11)
\end{aligned}$$

As described in above grammar, humans usually have two symmetrical arms and two symmetrical legs.

Motor coordination grammar $\mathcal{G}^{crd}$. Motor coordination is the ability to coordinate muscle synergies or movement primitives [74], which characterizes the coordinated involvement of different body parts in different actions. Studies [75], [76] show that there exist mutual couplings between the involved components and it can be used to reduce the degree of freedom in the model. Here we mainly consider simplified left-right motor coordination, *i.e.*, the rhythmic alternating left and right limb movement, during the locomotion. Motor coordination grammar is complementary to the symmetry grammar in describing the relations between upper body and lower body.

As illustrated in Figure 1 the green skeleton, we define two motor coordination grammar to represent constraints on people coordinated movements:

$$\begin{aligned}
\mathcal{G}^{crd}_{lr} &: \mathcal{G}^{kin}_{l.arm} \leftrightarrow \mathcal{G}^{kin}_{r.leg}\ , && (12)\\
\mathcal{G}^{crd}_{rl} &: \mathcal{G}^{kin}_{r.arm} \leftrightarrow \mathcal{G}^{kin}_{l.leg}\ . && (13)
\end{aligned}$$

In this paper, we extend Bi-directional RNN [77] (BRNN) to represent our pose grammar, which formulates the sequential data into a hidden Markov Chain model and describes relations among neighboring nodes with sharing weights. Unlike standard RNN which can only work in one direction, BRNN encodes messages passing from two directions, *i.e.*, forward and backward, which simulates the undirected relations between two neighboring joints in the pose grammar.

For a node t in the grammar, we use $t-1$ and $t+1$ to denote the previous node and the next node, respectively. For example, in Equation (5), node representing *thorax* is neighbored with the previous node for *neck* and the next node for *spine*. Given an input feature encoding a_t for node t, the output y_t is jointly determined by the bi-directional states h^f_t and h^b_t:

$$y_t = \phi(W^f_y h^f_t + W^b_y h^b_t + b_y), \tag{14}$$

where $\phi(\cdot)$ is the softmax function, h^f_t and h^b_t the forward and backward hidden states, respectively. h^f_t and h^b_t are computed recursively, that is,

$$\begin{aligned}
h^f_t &= \tanh(W^f_h h^f_{t-1} + W^f_a a_t + b^f_h)\ ,\\
h^b_t &= \tanh(W^b_h h^b_{t+1} + W^b_a a_t + b^b_h)\ ,
\end{aligned} \tag{15}$$

where W indicates the linear weights and b denotes the bias term.

As shown in Figure 2, we build a two-layer hierarchy of BRNNs for modeling our three kinds of pose grammar, where all BRNNs shares the same cell size 1024 (*i.e.*, 512 cells passing forward and 512 cells passing backward) and the same equation in Equation (14). The three pose grammar is represented by the edges between BRNNs nodes or implicitly encoded into BRNN architecture.

For the bottom layer, five BRNNs are built for modeling the five relations defined in kinematics grammar. More specifically, they accept the pose-aligned features from our base network as input, and generate estimation for a 3D joint at each time step. The information is forward/backward propagated efficiently over the two states with BRNN, thus the five Kinematics relations are implicitly modeled by the bi-directional chain structure of corresponding BRNN. Note that we take the advantages of recurrent natures of RNN for capturing our chain-like grammar, instead of using RNN for modeling the temporal dependency of sequential data.

For the top layer, four BRNN nodes are derived in total, two for symmetry relations and two for motor coordination relations. For the symmetry BRNN nodes, taking $\mathcal{G}^{sym}_{arm}$ node as an example, it takes the concatenated 3D-joints (totally 6 joints) from the $\mathcal{G}^{kin}_{l.arm}$ and $\mathcal{G}^{kin}_{r.arm}$ BRNNs in the bottom layer in all times as inputs, and produces estimations for the six 3D-joints taking their symmetry relations into account. Similarly, for the coordination nodes, such as $\mathcal{G}^{crd}_{l\to r}$, it leverages the estimations from $\mathcal{G}^{kin}_{l.arm}$ and $\mathcal{G}^{kin}_{r.leg}$ BRNNs and refines the 3D joint estimations according to motor coordination grammar.

Note a joint could appear multiple times in the pose grammar. We apply mean-pooling on outputs from all nodes in all BRNNs representing this joint and regard it as the jointly optimized configuration for this joint.

In this way, we organize three kinds of human pose grammar as a hierarchical BRNN model and the outputs from the top layer is considered as the final 3D pose estimations.

5 Learning

In this section, we first discuss the loss functions and learning algorithm overview and then propose a data augmentation technique to enhance the model robustness and generalization ability.

5.1 Loss Function

We first construct the training set Ω to train our model,

$$\Omega = \{(\mathbf{I}^k, \hat{\mathbf{U}}^k, \hat{\mathbf{V}}^k) : k = 1, \dots, N_\Omega\}, \tag{16}$$

$$\hat{\mathbf{U}}^k = \{\hat{u}_i^k : i = 1, \dots, N_v\}, \tag{17}$$

$$\hat{\mathbf{V}}^k = \{\hat{v}_i^k : i = 1, \dots, N_v\}, \tag{18}$$

where $\hat{\mathbf{U}}^k$ and $\hat{\mathbf{V}}^k$ denote ground-truth 2D and 3D poses for the k-th training sample, respectively. The mapping function $f(\mathbf{I}, \mathbf{U}; \theta)$ is learned via Maximum Likelihood Estimation (MLE), that is,

$$\begin{aligned}\theta^* &= \arg\max_\theta \mathcal{L}(\theta; \Omega) = \arg\min_\theta \ell(\theta; \Omega)\\ &= \arg\min_\theta \frac{1}{N_\Omega} \sum_{k=1}^{N_\Omega} \ell(\theta; \mathbf{I}^k, \hat{\mathbf{U}}^k, \hat{\mathbf{V}}^k).\end{aligned} \tag{19}$$

In this paper, we define the loss function $\ell(\theta; \mathbf{I}^k, \hat{\mathbf{U}}^k, \hat{\mathbf{V}}^k)$ to penalize both joint-wise and bone-wise errors between predicted 3D pose and true 3D pose, that is,

$$\ell(\theta; \mathbf{I}^k, \hat{\mathbf{U}}^k, \hat{\mathbf{V}}^k) = \sum_{i=1}^{N_v} \frac{\|v_i^k - \hat{v}_i^k\|_2}{N_v} + \sum_{(i,j)\in E} \frac{\|e_{ij}^k - \hat{e}_{ij}^k\|_2}{N_e}, \tag{20}$$

where $e_{ij} = v_i - v_j$ denotes the bone between joint i and j, E denotes the set of bones in the skeleton and N_e denotes the number of bones. Instead of considering all pairwise relations among joints, we only regard physically-connected neighboring joints as bones, as defined in the kinematics grammar $\mathcal{G}^{kin}$. The first loss term enforces the absolute locations at each joint to be accurate and the second loss term penalizes the bone displacement in the human skeleton, similar to [37].

5.2 Algorithm

The model parameters θ that need to be learned include weights from the geometry feature encoder, the appearance feature encoder and the 3D pose grammar network. Since the whole model could easily overfit with a simple end-to-end training, we divide the entire learning process into three phases:

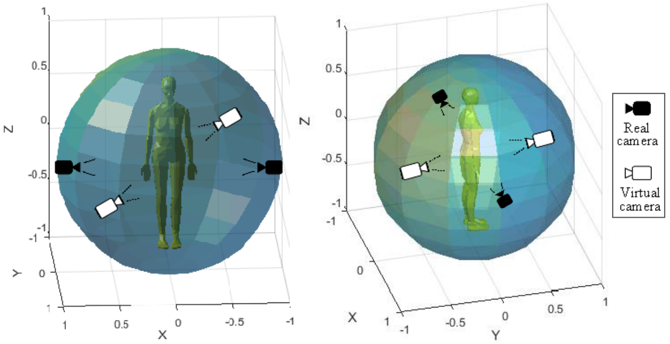


Fig. 3. Illustration of virtual camera simulation. The black camera icons stand for real camera settings while the white camera icons simulated virtual camera settings.

i) Learning geometry feature encoder from Mocap datasets using ground-truth 2D-3D pose pairs with loss function defined in Equation (19). Weights for other modules are frozen and input RGB images are not needed in this phase. The training set is augmented with virtual 2D poses by projecting each 3D pose into 2D poses under different camera viewpoints, as elaborated in Section 5.3.1.

ii) Learning appearance feature encoder from both MoCap and in-the-wild datasets with loss function defined in Equation (19). Weights for other modules are frozen as the first phase. We propose a label propagation algorithm for augmenting unpaired in-the-wild samples and train our appearance model on hybrid datasets to improve model robustness and mitigate overfitting, as discussed in Section 5.3.2.

iii) Attaching pose grammar network on the top of the trained base network, and fine-tuning the whole network in an end-to-end manner with Mocap datasets and augmented data from the first two phases.

5.3 Data Augmentation

We develop two data augmentation techniques to improve the model robustness and generalization ability for the corresponding learning phases.

5.3.1 2D Pose Augmentation

We conduct an empirical study on popular 3D pose estimation datasets (*e.g.*, *Human3.6M*, *HumanEva*) and notice that there are usually limited number of cameras (3-4 on average) recording the human subject. This raises the doubt whether learning on such dataset can lead to a generalized 3D pose estimator applicable in other scenes with different camera positions. We believe that augmenting 2D poses from unseen camera views will improve the model performance and generalization ability. For this, we propose a novel **Pose Sample Simulator** (PSS) to generate additional training samples. The generation process consists of two steps: i) projecting ground-truth 3D pose $\hat{\mathbf{V}}$ onto virtual camera planes to obtain ground-truth 2D pose $\hat{\mathbf{U}}$, ii) simulating 2D pose detections $\mathbf{U}$ by sampling conditional probability distribution $p(\mathbf{U}|\hat{\mathbf{U}})$.

In the first step, we first specify a series of virtual camera calibrations. Namely, a virtual camera calibration is specified by quoting intrinsic parameters K' from other real cameras and simulating reasonable extrinsic parameters (*i.e.*, camera locations T' and orientations R'). As illustrated in Figure 3, two white virtual camera calibrations are determined by the other two real cameras. Given a specified virtual camera, we can perform a perspective projection of a ground-truth 3D pose $\hat{\mathbf{V}}$ onto the virtual camera plane and obtain the corresponding ground-truth 2D pose $\hat{\mathbf{U}}$.

In the second step, we first model the conditional probability distribution $p(\mathbf{U}|\hat{\mathbf{U}})$ to mitigate the discrepancy between 2D pose detections $\mathbf{U}$ and 2D pose ground-truth $\hat{\mathbf{U}}$. Assuming $p(\mathbf{U}|\hat{\mathbf{U}})$ follows a mixture of Gaussian distribution, that is,

$$p(\mathbf{U}|\hat{\mathbf{U}}) = p(\epsilon) = \sum\nolimits_{j=1}^{N_G} \omega_j \, \mathbb{N}(\epsilon; \mu_j, \Sigma_j), \tag{21}$$

where $\epsilon = \mathbf{U} - \hat{\mathbf{U}}$, N_G denotes the number of Gaussian distributions, ω_j denotes a combination weight for the j-th component, $\mathbb{N}(\epsilon; \mu_j, \Sigma_j)$ denotes the j-th multivariate Gaussian distribution with mean μ_j and covariance Σ_j. As suggested in [2], we set $N_G = 42$. For efficiency issues, the covariance matrix Σ_j is assumed to be in the form:

$$\Sigma_j = \begin{bmatrix} \sigma_{j,1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_{j,i} \end{bmatrix}, \quad \sigma_{j,i} \in \mathbb{R}^{2\times 2} \tag{22}$$

where $\sigma_{j,i}$ is the covariance matrix for joint u_i at j-th multivariate Gaussian distribution. This constraint enforces independence among each joint u_i in 2D pose $\mathbf{U}$.

The probability distribution $p(\mathbf{U}|\hat{\mathbf{U}})$ can be efficiently learned using an EM algorithm, with E-step estimating combination weights ω and M-step updating Gaussian parameters μ and Σ. We utilizes K-means clustering to initialize parameters as a warm start. The learned mean μ_j of each Gaussian can be considered as an atomic pose representing a group of similar 2D poses. We visualize some atomic poses in Figure 4.

Given a 2D pose ground-truth $\hat{\mathbf{U}}$, we sample $p(\mathbf{U}|\hat{\mathbf{U}})$ to generate simulated detections $\mathbf{U}$, which reduces the discrepancy between training and testing data. Only 2D and 3D pose pairs are needed in the first learning phase, while black masks and simulated detections are combined together in the third learning phase for end-to-end joint training of the whole framework. The appearance feature encoder is essentially frozen for the augmented training samples, as it is not a trivial task to synthesize realistic images from arbitrarily unseen camera views.

5.3.2 Image Augmentation

To train the appearance feature encoder, we need 2D images with 2D and 3D pose annotations which are only available in constrained environments with motion capture systems. The appearance model is likely to overfit with training on those images alone, due to poor generalization to complex scenarios in the wild.

We propose a label propagation algorithm for unpaired in-the-wild images (*e.g.*, LSP [63], MPII [64]), using paired 2D and 3D poses from Mocap datasets and the trained geometry feature encoder from the first phase. The proposed label propagation algorithm consists of two steps: i) matching unpaired 2D poses with paired 2D poses from Mocap datasets, ii) verifying matched candidates using the early-stage model.

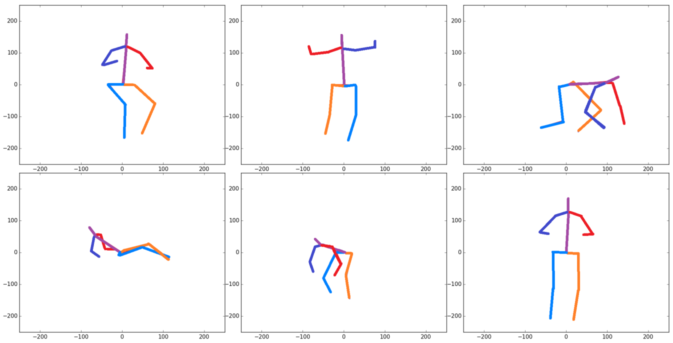


Fig. 4. Examples of learned 2D atomic poses in probability distribution $p(\mathbf{U}|\hat{\mathbf{U}})$.

First, we match each 2D pose of in-the-wild images with 2D poses in the MoCap dataset and return a set of 3D poses associated with the matched 2D poses. We rotate 2D poses with different angles because the diverse poses (*e.g.*, doing sports) in common pose datasets can hardly be observed under constrained conditions in MoCap datasets. We measure the matching of 2D poses by the mean distance per joints and keep top 10 matched 3D poses with distances below a certain threshold as candidates for the next step.

Second, we apply the trained geometry feature encoder on the 2D pose ground-truth to obtain an initial estimation of the 3D pose. We further compare the kept 3D pose candidates with the estimated 3D pose and pick the one with the minimum distance as the ground-truth 3D pose for this sample. Namely, the label from a paired sample (*i.e.*, in-the-wild data) is propagated to a unpaired sample (*i.e.*, Mocap data). We then browser through all the obtained 3D poses to filter out those bad matches.

Augmented image samples are first used to train the appearance feature encoder alone in the second learning phase and then used to augment the training set Ω in the third joint learning phase.

6 Experiments

In this section, we first introduce datasets and settings for evaluation, and then report our results and comparisons with state-of-the-art methods, and finally conduct an ablation study on components in our method.

6.1 Datasets

We evaluate our method quantitatively and qualitatively on four popular 3D pose estimation datasets.

Human3.6M [90], [1] is the current largest dataset for human 3D pose estimation, which consists of 3.6 million 3D human poses and corresponding video frames recorded from 4 different cameras. Cameras are located at the front, back, left and right of the recorded subject, with around 5 meters away and 1.5 meter height. In this dataset, there are 11 actors in total and 15 different actions performed (*e.g.*,

TABLE 1
Quantitative comparisons of Average Euclidean Distance (in mm) between the estimated pose and the ground-truth on *Human3.6M* under *Protocol #1*, *Protocol #2*, *Protocol #3* and the proposed *Protocol #4/4**. '-' indicates that the results were not reported for the respective action class in the original paper. Lower values are better. The best score is marked in **bold**.

Protocol #1	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Tekin *et al.*(CVPR'16) [78]	102.4	147.2	88.8	125.3	118.0	182.7	112.4	129.2	138.9	224.9	118.4	138.8	126.3	55.1	65.8	125.0
Zhou *et al.*(CVPR'16) [17]	87.3	109.3	87.0	103.1	116.1	143.3	106.8	99.7	124.5	199.2	107.4	118.0	114.2	79.3	97.7	113.0
Du *et al.*(ECCV'16) [79]	85.1	112.7	104.9	122.1	139.1	135.9	105.9	166.2	117.5	226.9	120.0	117.7	137.4	99.3	106.5	126.5
Sanzari *et al.*(ECCV'16) [80]	48.8	56.3	95.9	84.7	96.4	105.5	66.3	107.4	116.8	129.6	97.8	65.9	130.4	92.5	102.2	93.1
Chen *et al.*(CVPR'17) [81]	89.9	97.6	89.9	107.9	107.3	139.2	93.6	136.0	133.1	240.1	106.6	106.2	87.0	114.0	90.5	114.1
Pavlakos *et al.*(CVPR'17) [31]	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Moreno *et al.*(CVPR'17) [82]	69.5	80.1	78.2	87.0	100.7	76.0	69.6	104.7	113.9	89.6	102.7	98.4	79.1	82.4	77.1	87.3
Tome *et al.*(CVPR'17) [83]	64.9	73.4	76.8	86.4	86.2	110.6	68.9	74.7	110.1	173.9	84.9	85.7	86.2	71.3	73.1	88.3
Tekin *et al.*(ICCV'17) [84]	85.0	108.7	84.3	98.9	119.3	95.6	98.4	93.7	73.7	170.4	85.0	116.9	113.7	62.0	94.8	100.0
Zhou *et al.*(ICCV'17) [33]	54.8	60.7	58.2	71.4	62.0	65.5	53.8	55.6	75.2	111.6	64.1	66.0	51.4	63.2	55.3	64.9
Martinez *et al.*(ICCV'17) [6]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Yang *et al.*(CVPR'18) [23]	51.5	58.9	**50.4**	57.0	62.1	65.4	49.8	52.7	69.2	85.2	**57.4**	58.4	43.6	60.1	47.7	58.6
Hossain *et al.*(ECCV'18) [25]	48.4	**50.7**	57.2	55.2	63.1	72.6	53.0	**51.7**	**66.1**	80.9	59.0	57.3	62.4	46.6	49.6	58.3
Habibie *et al.*(CVPR'19) [36]	54.0	65.1	58.5	62.9	67.9	75.0	54.0	60.6	82.7	98.2	63.3	61.2	66.9	50.0	56.5	65.7
Zhao *et al.*(CVPR'19) [37]	47.3	60.7	51.4	60.5	**61.1**	**49.9**	**47.3**	68.1	86.2	**55.0**	67.8	61.0	**42.1**	60.6	**45.3**	**57.6**
Ours (ICCV'17) [9]	90.1	88.2	85.7	95.6	103.9	92.4	90.4	117.9	136.4	98.5	103.0	94.4	86.0	90.6	89.5	97.5
Ours (AAAI'18) [10]	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
Ours Full	**47.1**	52.8	54.2	**54.9**	63.8	72.5	51.7	54.3	70.9	85.0	58.7	**54.9**	59.7	**43.8**	47.1	58.1
Protocol #2	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Akhter *et al.*(CVPR'15) [85]	199.2	177.6	161.8	197.8	176.2	195.4	167.3	160.7	173.7	177.8	186.5	181.9	198.6	176.2	192.7	181.5
Zhou *et al.*(PAMI'16) [86]	99.7	95.8	87.9	116.8	108.3	93.5	95.3	109.1	137.5	106.0	107.3	102.2	110.4	106.5	115.2	106.1
Bogo *et al.*(ECCV'16) [87]	62.0	60.2	67.8	76.5	92.1	77.0	73.0	75.3	100.3	137.3	83.4	77.3	86.8	79.7	87.7	82.3
Sanzari *et al.*(ECCV'16) [80]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	93.1
Moreno *et al.*(CVPR'17) [82]	66.1	61.7	84.5	73.7	65.2	67.2	60.9	67.3	103.5	74.6	92.6	69.6	71.5	78.0	73.2	74.0
Pavlakos *et al.*(CVPR'17) [31]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	51.9
Tome *et al.*(CVPR'17) [83]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	79.6
Tekin *et al.*(ICCV'17) [84]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	50.1
Martinez *et al.*(ICCV'17) [6]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Yang *et al.*(CVPR'18) [23]	**26.9**	**30.9**	**36.3**	**39.9**	**43.9**	**47.4**	**28.8**	**29.4**	**36.9**	**58.4**	**41.5**	**30.5**	**29.5**	42.5	**32.2**	**37.7**
Hossain *et al.*(ECCV'18) [25]	35.7	39.3	44.6	43.0	47.2	54.0	38.3	37.5	51.6	61.3	46.5	41.4	47.3	**34.2**	39.4	44.1
Habibie *et al.*(CVPR'19) [36]	43.7	46.9	45.4	48.0	50.2	54.6	40.6	41.6	60.7	75.6	48.8	46.8	47.5	36.9	43.9	49.2
Ours (ICCV'17) [9]	72.5	69.9	69.2	78.3	80.0	71.7	70.8	83.1	105.7	76.0	83.5	76.4	69.0	75.2	79.6	77.4
Ours (AAAI'18) [10]	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	45.7
Ours Full	36.7	39.5	41.5	42.6	46.9	53.5	38.2	36.5	52.1	61.5	45.0	42.7	45.2	35.3	40.2	43.8
Protocol #3	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Kostrikov *et al.*(BMVC'14) [88]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	115.7
Rogez *et al.*(NIPS'14) [89]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	88.1
Yasin *et al.*(CVPR'16) [5]	88.4	72.5	108.5	110.2	97.1	81.6	107.2	119.0	170.8	108.2	142.5	86.9	92.1	165.7	102.0	110.2
Tome *et al.*(CVPR'17) [83]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	70.7
Ours (ICCV'17) [9]	62.8	69.2	79.6	78.8	80.8	72.5	73.9	96.1	106.9	88.0	86.9	70.7	71.9	76.5	73.2	79.5
Ours (AAAI'18) [10]	32.9	43.7	52.1	46.1	51.0	57.1	43.3	46.5	60.1	73.2	52.1	42.6	50.7	38.6	40.8	48.7
Ours Full	**31.8**	**42.0**	**49.9**	**44.6**	**48.9**	**54.7**	**42.2**	**45.1**	**58.2**	**71.3**	**50.6**	**41.2**	**48.9**	**37.2**	**39.3**	**47.1**
Protocol #4	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Pavlakos *et al.*(CVPR'17) [31]	79.2	85.2	78.3	89.9	86.3	87.9	75.8	81.8	106.4	137.6	86.2	92.3	72.9	82.3	77.5	88.6
Zhou *et al.*(ICCV'17) [33]	61.4	70.7	**62.2**	76.9	71.0	**81.2**	67.3	71.6	96.7	126.1	68.1	76.7	**63.3**	72.1	68.9	75.6
Martinez *et al.*(ICCV'17) [6]	65.7	68.8	92.6	79.9	84.5	100.4	72.3	88.2	109.5	130.8	76.9	81.4	85.5	69.1	68.2	84.9
Ours (ICCV'17) [9]	103.9	103.6	101.1	111.0	118.6	105.2	105.1	133.5	150.9	113.5	117.7	108.1	100.3	103.8	104.4	112.1
Ours (AAAI'18) [10]	**57.5**	**57.8**	81.6	68.8	**75.1**	85.8	**61.6**	**70.4**	95.8	106.9	**68.5**	70.4	73.8	58.5	59.6	72.8
Ours Full	59.5	65.6	66.7	**65.7**	78.3	72.2	64.6	71.3	**89.3**	**105.4**	71.9	**64.6**	**64.0**	**52.2**	**57.4**	**69.9**
Protocol #4*	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Pavlakos *et al.*(CVPR'17) [31]	76.9	84.3	74.0	88.0	90.0	88.7	74.8	77.4	107.2	140.6	88.4	87.7	69.9	73.4	74.2	86.4
Zhou *et al.*(ICCV'17) [33]	58.7	70.1	**57.8**	75.7	74.2	82.8	66.1	68.3	95.8	128.3	69.5	72.7	**60.7**	64.6	65.5	74.1
Martinez *et al.*(ICCV'17) [6]	63.8	68.4	86.0	78.2	88.2	102.0	71.8	85.5	109.7	134.8	78.9	75.2	80.3	62.2	65.1	83.3
Ours (ICCV'17) [9]	100.7	102.5	96.7	108.6	126.9	104.6	105.0	126.8	152.7	117.0	118.2	100.6	93.7	93.1	101.2	112.0
Ours (AAAI'18) [10]	59.2	**60.6**	73.1	67.6	**67.2**	80.9	62.1	67.0	86.6	100.0	63.8	71.1	77.6	61.5	61.1	70.6
Ours Full LOO	**57.2**	61.5	64.4	**64.6**	71.0	**76.9**	**62.0**	**63.5**	**84.8**	**101.1**	**65.8**	**65.1**	70.5	**52.9**	**57.4**	**67.9**
Ours Full LO-View0	59.5	65.6	66.7	65.7	78.3	72.2	64.6	71.3	89.3	105.4	71.9	64.6	64.0	52.2	57.4	69.9
Ours Full LO-View1	55.9	61.3	58.1	64.8	68.2	81.0	61.2	58.3	81.4	91.4	62.2	64.7	75.2	50.4	58.3	66.2
Ours Full LO-View2	54.8	58.9	60.7	61.3	71.2	74.7	60.7	58.5	83.2	109.4	66.1	60.6	66.5	48.6	53.5	65.9
Ours Full LO-View3	58.4	60.2	72.2	66.6	66.5	79.5	61.4	65.9	85.4	98.3	62.9	70.4	76.3	60.5	60.5	69.7

greeting, eating and walking, *etc*.). The 3D pose ground-truth is captured by a motion capture (Mocap) system and all camera parameters (intrinsic and extrinsic parameters) are provided. This dataset is captured in a controlled indoor environment.

HumanEva [7] is another widely used dataset for human 3D pose estimation, which is also collected in a controlled indoor environment using a Mocap system. In this paper, we use *HumanEva-I* dataset for experimental validation, which consists of 4 subjects performing 6 predefined actions (*e.g.*, jog, gesture and combo) under 7 cameras (3 color cameras and 4 grayscale cameras).

Human-Human-Object Interaction Dataset (HHOI) [8] is a newly released dataset that contains 3D human interactions captured by MS Kinect v2 sensor. It includes 3 types of human-human interactions: shake hands, high five and pull up and 2 types of human-object-human interactions: throw and catch, hand over a cup. There are 8 actors performing 23.6 instances per interaction on average. The data is collected in a common office with clutter background.

MPII [2] is a challenging benchmark for 2D human pose estimation in the wild, containing a large amount of human images in the wild. We only validate our method on this dataset qualitatively since no 3D pose ground-truth is provided.

6.2 Evaluation Protocols

6.2.1 Human3.6M

For *Human3.6M* [1], there are mainly three evaluation protocols used to measure the performance.

Protocol #1, the most standard evaluation protocol on *Human3.6M*, is widely used in the literature [1], [29], [17], [84]. The original frame rate of 50fps is down-sampled to 10fps and the evaluation is on sequences coming from all 4 cameras and all trials. The first 5 subjects (S1, S5, S6, S7 and S8) are used for training and the last 2 subjects (S9 and S11) are used for testing. The reported 3D error metric is computed as the Euclidean distance from the estimated 3D joints to the ground-truth, averaged over all 17 joints of the *Human3.6M* skeletal model.

Protocol #2, followed by [87], [85], [91], [86], selects the same subjects for training and testing as *Protocol #1*. However, evaluation is only on sequences captured from the frontal camera ("cam 3") from trial 1 and all the frames are used. The predictions are post-processed via a rigid transformation (*i.e.*each estimated 3D pose is aligned with the ground-truth pose, on a per-frame basis, using Procrustes analysis) before comparing to the ground-truth.

Protocol #3 was used in [88], [5], [89], [83]. The training set consists of 6 subjects (S1, S5, S6, S7, S8 and S9) while the testing set only contains 1 subject (S11). The testing data is sub-sampled from S11 with an interval of 64 frames. The evaluation is performed on sequences from all 4 cameras and all trials. Some poses without synchronized images are omitted and the total testing set has 3,612 poses. Similar to *Protocol #2*, the estimated skeleton was first aligned to the ground-truth one by Procrustes transformation before measuring the joint distances.

In above three protocols, the same 4 camera views are both used for training and testing. This raises the question whether the learned estimator over-fits to training camera parameters. To validate the generalization ability of different models, we propose a new protocol based on different camera view partitions for training and testing.

Proposed Protocol #4 and #4*. In [10], we propose to use subjects S1, S5, S6, S7 and S8 in the first 3 camera views for training and subjects S9 and S11 in the last camera view for testing, which is referred as *Protocol #4*. In this paper, we further extend the cross-view training/testing data partition in *Protocol #4* to a Leave-One-Out setting. Specifically, we keep the same subject partitions for the training and testing set, but conduct four experiments on four corresponding cross-view training/testing data partitions, i.e., {*view0, view1, view2*}/{*view3*}, {*view0, view1, view3*}/{*view2*}, {*view0, view2, view3*}/{*view1*}, {*view1, view2, view3*}/{*view0*}. The mean errors are computed over all four experiment settings. The new protocol is referred as *Protocol #4**. This Leave-One-Out cross validation scheme is more robust compared with the original fixed camera view partition. Compared with other evaluation protocols, the suggested protocols guarantee that not only subjects but also camera views are different for training and testing, eliminating over-fitting of subject appearance and camera parameters, respectively.

6.2.2 HumanEva

For *HumanEva-I* [2] dataset, we follow the standard protocol described in [92], [88], [5], [31] for a fair comparison. The training sequences are used for training and the validation sequences for evaluation. The performance is evaluated on the jogging and walking sequences from 3 subjects (S1, S2, S3) and the first RGB camera. A rigid transformation is performed before computing the mean reconstruction error.

6.2.3 HHOI

To evaluate how our method can be generalized to data from a totally different environment, we train model on *Human3.6M* dataset and test it on *HHOI* dataset which is captured with Kinect sensor in a casual environment. We pick 13 joints defined by Kinect and also use mean per joint error as the evaluation metric. Each action instance is down-sampled at 10fps for efficient computation and both persons in each action are evaluated. We still use the focal length from *Human3.6M* to recover 3D poses and the poses are compared up to a rigid transformation and also scale transformation.

6.3 Implementation Details

We implement our method using Tensorflow and Keras as back-end. We first train our base network for 200 epoch. In the first step, the learning rate is set as 10^{-4}. We use all training samples from each dataset and all images from MPII-LSP-extended dataset [62] for the joint training of the appearance feature encoder. In the second step, the learning rate is set as 0.001 with an exponential decay rate 0.96 and decay step 100000 and the dropout rate is set as 0.5. In the third step, the learning rate is set as 10^{-5} to guarantee model stability in the final training phase. The batch size is set to 64 and we adopt Adam optimizer for the mini-batch gradient descent in all steps.

TABLE 2
Quantitative comparisons of the mean reconstruction error (mm) on *HumanEva-I*. The best score is marked in **bold**.

Method	Walking (Act 2, Cam 1)			Jogging (Act 1, Cam 1)			Avg.
	S1	S2	S3	S1	S2	S3	
Taylor *et al.*(CVPR'10) [93]	48.8	47.4	49.8	75.4	-	-	-
Bo *et al.*(IJCV'10) [94]	46.4	30.3	64.9	64.5	48.0	38.2	48.7
Sigal *et al.*(IJCV'12) [95]	66.0	69.0	-	-	-	-	-
Varun *et al.*(ECCV'12) [91]	161.8	182.0	188.6	-	-	-	-
Simo-Serra *et al.*(CVPR'12) [92]	99.6	108.3	127.4	109.2	93.1	115.8	108.9
Simo-Serra *et al.*(CVPR'13) [96]	65.1	48.6	73.5	74.2	46.6	32.2	56.7
Radwan *et al.*(ICCV'13) [97]	75.1	99.8	93.8	79.2	89.8	99.4	89.5
Wang *et al.*(CVPR'14) [98]	71.9	75.7	85.3	62.6	77.7	54.4	71.2
Belagiannis *et al.*(CVPR'14) [99]	68.3	-	-	-	-	-	-
Kostrikov *et al.*(BMVC'14) [88]	44.0	30.9	41.7	57.2	35.0	33.3	40.3
Elhayek *et al.*(CVPR'15) [100]	66.5	-	-	-	-	-	-
Akhter *et al.*(CVPR'15) [85]	186.1	197.8	209.4	-	-	-	-
Yasin *et al.*(CVPR'16) [5]	35.8	32.4	41.6	46.6	41.4	35.4	38.9
Tekin *et al.*(CVPR'16) [78]	37.5	25.1	49.2	-	-	-	-
Zhou *et al.*(CVPR'16) [17]	34.2	30.9	49.1	-	-	-	-
Zhou *et al.*(PAMI'16) [86]	100.0	98.9	123.1	-	-	-	-
Bogo *et al.*(ECCV'16) [87]	73.3	59.0	99.4	-	-	-	-
Noguer *et al.*(CVPR'17) [82]	19.7	**13.0**	**24.9**	39.7	20.0	21.0	26.9
Pavlakos *et al.*(CVPR'17) [31]	22.3	19.5	29.7	28.9	21.9	23.8	24.3
Tekin *et al.*(ICCV'17) [84]	27.2	14.2	31.7	-	-	-	-
Martinez *et al.*(ICCV'17) [6]	19.7	17.4	46.8	26.9	18.2	18.6	24.6
Hossain *et al.*(ECCV'18) [25]	19.1	13.6	43.9	**23.2**	**16.9**	**15.5**	22.0
Ours (ICCV'17) [9]	32.2	18.3	32.5	42.4	36.6	34.3	32.7
Ours (AAAI'18) [10]	19.4	16.8	37.4	30.4	17.6	16.3	22.9
Ours Full	**18.7**	16.0	35.7	28.9	**16.9**	15.6	**21.9**

We perform 2D pose detections using a state-of-the-art 2D pose estimator [4]. We fine-tuned the model on *Human3.6M* and use the pre-trained model on *HumanEva* and *MPII*. Our deep grammar network is trained with images and 2D pose detections as inputs and 3D pose ground-truth as outputs. For *Protocol #1*, *Protocol #2*, and *Protocol #3*, the data augmentation is omitted due to neglected performance change ($\pm 0.2mm$) and tripled training time. We refer the readers to Section 6.5.3 for more details and analysis on data augmentation. For *Protocol #4* and *Protocol #4**, in addition to the original 3 camera views, we further augment the training set with 6 virtual camera views on the same horizontal plane. Consider the circle which is centered at the human subject and locates all cameras is evenly segmented into 12 sectors with 30 degree angles each, and 4 cameras occupy 4 sectors. We generate training samples on 6 out of 8 unoccupied sectors and leave 2 closest to the testing camera unused to avoid overfitting. The 2D poses generated from virtual camera views are augmented by our PSS. During each epoch, we will sample our learned distribution once and generate a new batch of synthesized data.

Empirically, one forward and backward pass takes 35 ms on a Titan X GPU and a forward pass takes 20 ms only, allowing us to train and test our network efficiently.

6.4 Results and Comparisons

Human3.6M. We evaluate our method under all three protocols. We compare our method with 14 state-of-the-art methods [78], [79], [17], [81], [80], [87], [31], [91], [86], [85], [82], [33], [6], [23], [25], [36], [37] and report quantitative comparisons in Table 1. From the results, our method achieves state-of-the-art results across the vast majority of actions and obtains superior performance over other competing methods on average.

To verify our claims, we re-train three previous methods [31], [33], [6], which obtain top performance under *Protocol #1*, with the new dataset partitions under *Protocol #4* and *Protocol #4**. The quantitative results are reported in Table 1. The large performance drop of previous 2D-3D reconstruction models [31], [9], [33], [6], which demonstrates the blind spot of previous evaluation protocols and the over-fitting problem of those models.

Notably, our method outperforms previous methods in all evaluation protocols and shows a huge improvement under under cross-view evaluation in *Protocol #4* and *Protocol #4**. Additionally, the large performance gap of [6] under *Protocol #1* and *Protocol #4* ($62.9mm$ *vs* $84.9mm$) demonstrates that previous 2D-to-3D reconstruction networks easily over-fit to camera views. Our general improvements over different settings demonstrate our superior performance and good generalization.

HumanEva. We compare our method with 18 state-of-the-art methods [93], [94], [95], [91], [92], [96], [97], [98], [99], [88], [100], [85], [5], [86], [87], [82], [31], [6], [25]. The quantitative comparisons on *HumanEva-I* are reported in Table 2. As seen, our 3D pose estimator outperforms previous methods across the vast majority of subjects and on average.

HHOI. We implement a baseline 'Nearest' which matches the predicted 2D pose with 2D poses from *Human3.6M* and selects the depth from the 3D pose paired with the nearest 2D pose as the predicted depth. Note that the Kinect may produce unreasonable 3D poses because of occlusions and the evaluation with those poses cannot reflect true performance of compared methods, thus we go through each action video and select visually good sequences for quantitative comparisons. Specifically, we keep all videos from 'PullUp' and 'HandOver', and a few videos from 'HighFive' and 'ShakeHands'. We select the smaller error calculated among the predicted pose and its flipped one due to the left-right confusion of Kinect. The quantitative results are summarized in Table 3. The action 'PullUp' gets the biggest error among all actions due to the large pose variation.

For all three above datasets, we set up three baselines for our ICCV'17 framework [9], our AAAI'18 framework [10] and our new model. As reported in Table 1, Table 2 and Table 3, the new method outperforms our previous frameworks, which validates the effectiveness of incorporating appearance features.

MPII. We apply the pre-trained model from *Human3.6M* to natural images to see how well the learned model could be generalized to unseen images. We visualize sampled results generated by our method on *MPII* as well as *Human3.6M* in Figure 5. As seen, our method is able to accurately predict 3D pose for both indoor and in-the-wild images.

6.5 Ablation studies

In this section, we study different components of our model on *Human 3.6M* dataset under different protocols.

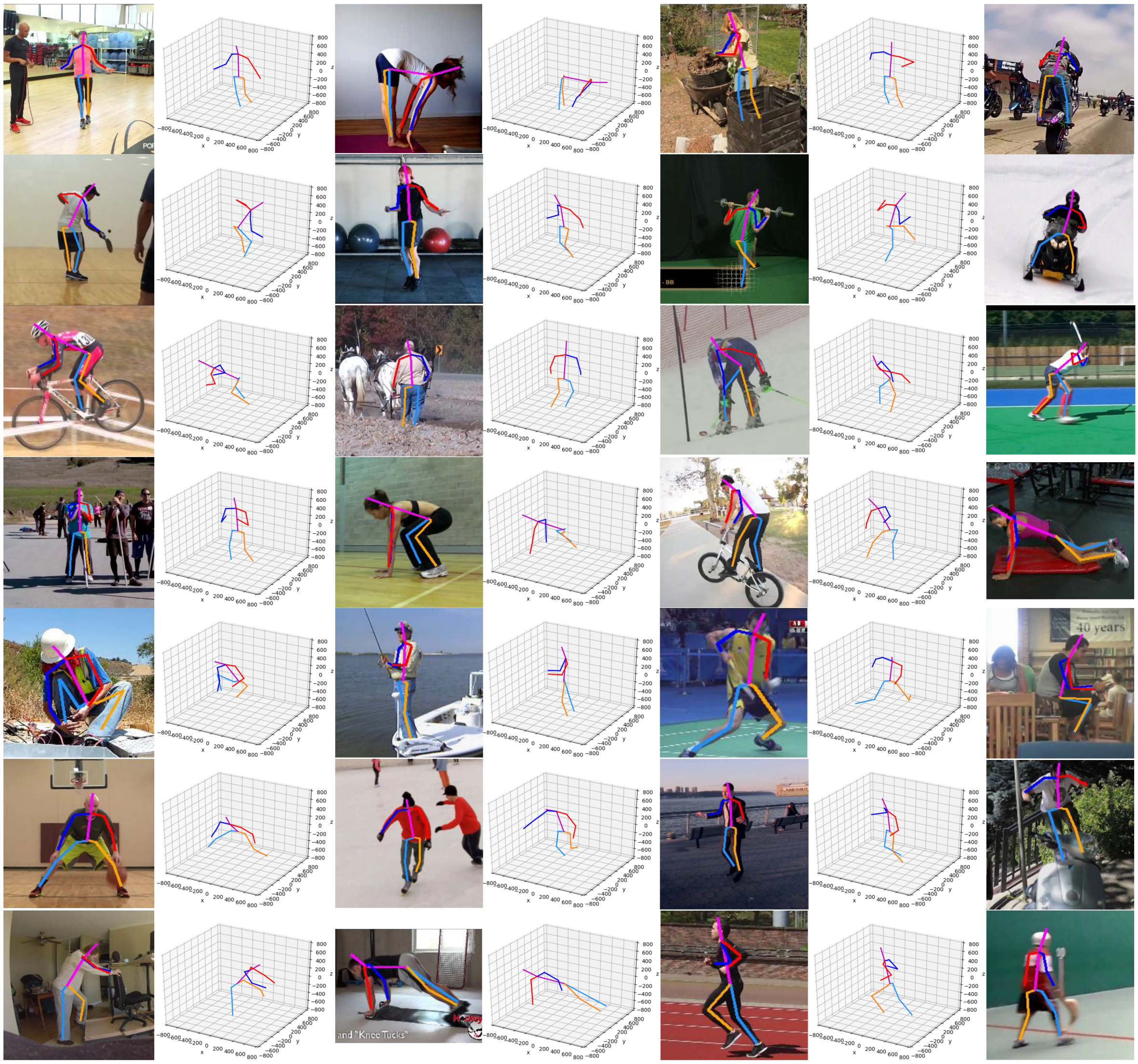


Fig. 5. Qualitative results of our method on *MPII* dataset. For each sample pair, we show the input image overlayed with the estimated 2D pose (left) and the estimated 3D pose from a novel view (right). Best viewed in color.

TABLE 3
Quantitative comparisons of Average Euclidean Distance (in mm) between the estimated pose and the ground-truth on *HHOI*. Lower values are better. The best score is marked in **bold**.

Method	PullUp	HandOver	HighFive	ShakeHands	Avg.
Nearest	161.2	126.2	117.3	129.6	133.6
Ours (ICCV'17) [9]	124.8	101.9	96.1	118.6	110.4
Ours (AAAI'18) [10]	60.9	63.3	61.5	72.1	64.5
Ours Full	**58.3**	**60.6**	**58.8**	**69.2**	**61.7**

TABLE 4
Ablation studies of features and loss terms on *Human3.6M* under *Protocol* #1, #2 and #4*. Numbers represent Average Euclidean Distance (in mm) between estimation and ground-truth and lower values are better. See text for detailed explanations.

Component	Variants	Proto #1	Proto #2	Proto #4*
Appearance	Ours w/o.	60.4	45.7	70.6
	Ours Full	58.1	43.8	67.9
Geometry	Ours w/o.	85.1	67.6	95.0
	1 block	85.1	67.6	74.6
	2 blocks (Ours Full)	58.1	43.8	67.9
	4 blocks	60.3	45.6	69.7
	8 blocks	61.0	46.0	70.3
Bone Loss	Ours w/o.	58.6	44.2	68.8
	Ours Full	58.1	43.8	67.9

6.5.1 *Effectiveness of Features and Loss Terms*

Features. We study the effectiveness of two feature encoders. As reported in Table 4, we observe the error increases by $\sim 4\%$ on all protocols when dropping off the appearance feature, . We then validate how the geometry feature works by either dropping it off (*i.e.,* appearance feature only) or changing to different number of building blocks. It can be observed adding geometry features can lead to huge performance improvements on all protocols and 2 blocks seem to be a fairly good choice in the proposed network architecture. This ablative study also indicates geometry features are more expressive than appearance

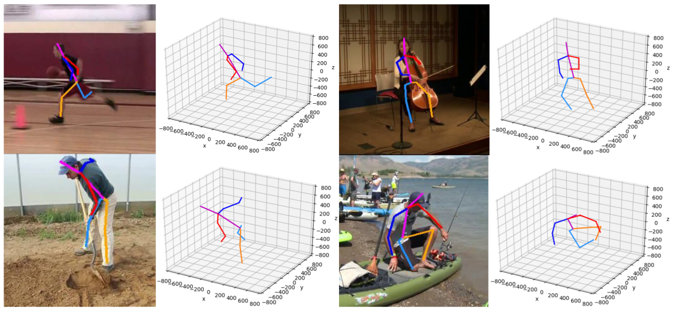


Fig. 6. Illustration of how symmetry constraints help pose estimation. Even when certain body parts are occluded/mis-detected in the input image, our model could still predict reasonable 3D pose configurations.

features in the task of 3D human pose estimation.

Loss Terms. We study the effectiveness of the newly added bone displacement loss. Comparing with the baseline which does not use the bone loss, we can observe that adding bone displacement constrains could lead to $\sim 1.5\%$ performance improvement across different evaluation protocols (see Table 4). Essentially, this loss serves as a binary loss on the bone 3D orientations, analogy to Part Affinity Field [101] used in 2d human pose estimation.

TABLE 5
Ablation studies of pose grammar on *Human3.6M* under *Protocol* #1, #2 and #4*. Numbers represent Average Euclidean Distance (in mm) between estimation and ground-truth. Lower values are better. See text for detailed explanations.

Methods	Components	Proto #1	Proto #2	Proto #4*
Ours (AAAI'18) [10] w/o. extra data	w/o.	69.4	53.8	92.0
	w. kin.	63.7	49.1	85.9
	w. kin.+sym.	61.8	46.8	83.4
	w. kin.+crd.	62.7	47.9	84.9
	w. all	60.4	45.7	82.6
Ours (AAAI'18) [10]	w/o.	65.2	50.3	75.1
	w. kin.	62.0	47.5	72.9
	w. kin.+sym.	60.9	46.2	71.4
	w. kin.+crd.	61.3	46.9	72.0
	w. all	60.3	45.6	70.6
Ours w/o. extra data	w/o.	67.3	52.4	88.1
	w. kin.	62.3	46.9	81.7
	w. kin.+sym.	59.7	44.8	79.5
	w. kin.+crd.	61.0	45.4	80.6
	w. all	58.1	43.8	78.5
Ours Full	w/o.	61.0	46.6	70.7
	w. kin.	59.3	45.1	68.5
	w. kin.+sym.	58.3	44.0	68.4
	w. kin.+crd.	58.7	44.5	68.8
	w. all	57.9	43.7	67.9

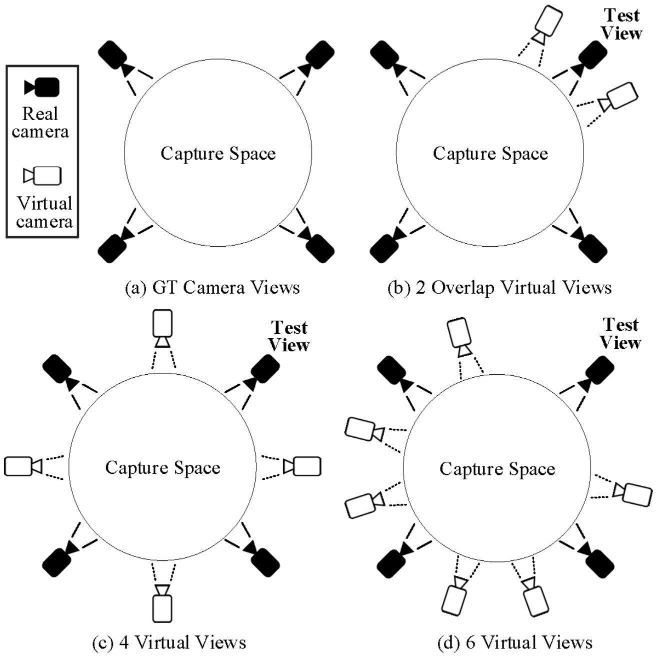
6.5.2 *Effectiveness of Pose Grammar*

We study the effectiveness of our pose grammar model, which injects high-level grammar constraints into our network. We first set up four baselines: 1) the base network without data augmentation from [10], 2) the base network with data augmentation from [10], 3) the proposed base network without data augmentation, and 4) the proposed base network with data augmentation. For all four methods, we examine the performance of different grammar rules: 1) without pose grammar, 2) kinematics grammar only, 3) kinematics and symmetry grammar, 4) kinematics and coordination, 5) all three grammar. As reported in Table 5, each pose grammar rule indeed helps constraint 3D pose configurations and improves the performance. Comparing the results among different methods, pose grammar shows good performance improvement on baselines (simple network architecture, no extra training data). When the model is fed with rich data, contribution of pose grammar decreases. Comparing results among all three pose grammar, we can also find that kinematics grammar plays a major rule in constraining the 3D human pose, then symmetry grammar, then the motor coordination grammar.

Noticed the marginally improvement reported in Fig. 5 might not fully illustrate how each grammar constrains the pose configuration, this could be potentially due to the relatively simple poses in *Human3.6M* dataset and the 3D supervisions dominates in training. We further dive deep to analyze how specifically each grammar contributes on challenging examples from in-the-wild datasets. As shown in Fig. 5, quantitative results on various challenging poses from *MPII* indicates our learned model generalizes well and generates reasonable 3D poses on the in-the-wild images, especially considering the challenging poses and heavy occlusions in those testing images, as well as lack of paired 3D groundtruth on the in-the-wild datasets for training.

Our vanilla model without grammar model represents each joint independently and relies on the neural network to implicitly learn the joint-wise relations, which could introduce undesired priors due to the data bias, e.g., most of the subjects in training set are in upright poses. *Kinematic grammar*, defined as chains among kinematically-connected body parts, helps reduce the state space between joints with parent-child relations. This is particularly beneficial for cases with partial occlusions, *e.g.*, a joint is occluded and its parent/child joint is visible. *Symmetry grammar*, together with *kinematic grammar*, plays a key role in handling challenging poses and occlusions. As shown in Fig. 6, when some body parts are occluded/mis-detected from the input image, our model still produces reasonable estimations for those occluded body parts. Empirically, our learned model could utilize i) the estimated locations of their parent/child joints and ii) the configuration from its symmetric body parts to infer positions of the occluded joints.

We conduct an ablative study on *MPII* dataset to check if the symmetrical body parts have similar lengths at the end of the pose estimation process. Specially, we compute the bone length differences between left and right limbs (*i.e.*, forearm, arm, thigh, leg) and average over all the *MPII* dataset. Estimated 3D poses from i) model without any pose grammar, ii) model with kinematic grammar only, iii) model with kinematic and symmetry grammar yield mean bone length differences of $13.3mm$, $11.7mm$ and $9.1mm$, respectively. This further validates kinematic and symmetry grammar enforce constraints to preserve the plausibility of human body pose configuration. It is worth noting even human performs poorly to provide the groundtruth for those challenging examples due to depth ambiguity.



(a) GT Camera Views (b) 2 Overlap Virtual Views (c) 4 Virtual Views (d) 6 Virtual Views

Fig. 7. Illustration of four different viewpoint augmentation strategies. See text for detailed explanations.

6.5.3 Effectiveness of Data Augmentation

Viewpoint Augmentation. We study different strategies for selecting virtual viewpoints. It worth noting that the proposed data augmentation method aims to improve the model robustness against unseen views. Yet it can hardly be evaluated fairly using existing 3D pose datasets, where they use very few cameras (less than 10) and with Mocap stage setup bias. As illustrated in Fig. 7, we conduct our study on *Human 3.6M* data and explore four viewpoint selection strategies:

a) Directly using the ground-truth test camera view. This does not introduce any redundant viewpoints and serves as the upper bound for data augmentation.

b) Simulating two camera views locating ±30 degrees next to the ground-truth camera view for testing. This serves as an overfitted viewpoint augmentation.

c) Simulating four camera views evenly distributed around all the cameras, where each virtual camera is 45 degree away from the two neighboring real cameras.

d) Simulating eight camera views, together with the 4 real cameras to segment the circle centering the human subject into 12 sectors with 30 degrees each. We ignore the two cameras which fall into ±30 degrees of the testing camera view to avoid overfitting and keep the rest six camera views.

As shown in Table 6, our strategy is inferior to selecting ground-truth and the overfitting setting by a small margin. Since we expect the viewpoint augmentation should be agnostic to the testing scenario, our augmentation strategy provides people a general sense about how much improvement we could expect from using our data augmentation. Even though performance could be better if we select a viewpoint close to the one for testing, such strategy violates our dataset assumption and cannot be generalized to other 3D pose datasets or in-the-wild images.

Sampling. We then evaluate the effects of our 2D pose sample simulator. As shown in Table 6, comparing the results of only using the data from original 3 camera views in *Human 3.6M* and the results of adding samples by generating ground-truth 2D-3D pairs from 6 extra camera views, we see error drops on all protocols, showing that extra training data indeed expand the generalization ability. Next, we compare our Pose Sample Simulator to a simple baseline, *i.e.*, generating samples by adding random noises to each joint, say an arbitrary Gaussian distribution or a white noise. Unsurprisingly, we observe error drops on all protocols yet it is worse than using the ground-truth 2D pose. This suggests that the conditional distribution $p(E|\hat{E})$ helps bridge the gap between detection results and ground-truth.

Augmentation for other methods. Finally, we re-train models proposed in [31], [33], [6] with the augmented data to validate when such data provides complementary information about 3D human pose in unseen views. As reported in Table 6, We can observe performance boosts for all these methods given the extra training data. This study also serves as a fair comparisons between our method and other competitive methods.

We also find that data augmentation from unseen views cannot benefit model performance much on protocol #1 and #2 for all methods and variants. This indicates that unseen views barely provide any complementary information for the model to learn 3D human body configuration if the training and testing camera views are highly overlapped. Due to the limitations of existing Mocap datasets, benefits from data augmentation could be underestimated. As shown in Fig. 5, qualitative results on in-the-wild images indicates our method generalizes well on unseen human poses and camera views.

TABLE 6
Ablation studies of viewpoint augmentation, pose sampling and augmentation for other methods on *Human3.6M* under *Protocol* #1, #2 and #4*. Numbers represent Average Euclidean Distance (in mm) between estimation and ground-truth. Lower values are better. See text for detailed explanations.

Components	Variants	Proto #1	Proto #2	Proto #4*
Virtual View	GT (upper bound)	57.5	43.4	65.7
	2 overlap views	57.9	43.6	66.7
	4 views	58.0	43.8	69.8
	6 views (Ours Full)	57.9	43.7	67.9
Sampling	w/o.	58.1	43.8	78.5
	w. GT	58.4	44.0	72.1
	w. simple	58.9	44.2	73.3
	w. PSS (Ours Full)	57.9	43.7	67.9
Methods	Components	Proto #1	Proto #2	Proto #4*
Pavlakos *et al.*(CVPR'17) [31]	w/o. extra data	71.9	51.9	86.4
	w. extra data	71.5	51.6	78.5
Zhou *et al.*(ICCV'17) [33]	w/o. extra data	64.9	-	74.1
	w. extra data	64.9	-	72.4
Martinez *et al.*(ICCV'17) [6]	w/o. extra data	62.9	47.7	83.3
	w. extra data	62.8	47.7	73.9
Ours Full	w/o. extra data	58.1	43.8	78.5
	w. extra data	57.9	43.7	67.9

7 Conclusions

In this paper, we propose a pose grammar model for the task of human 3D pose estimation, which encode appearance and geometry features of 2D human poses implicitly, and a set of knowledge over 3D human poses explicitly. The proposed pose grammar expresses the composition process of joints-part-pose following various principles (*i.e.*, kinematics, symmetry, and coordination). The network can be trained end-to-end with back-propagation. For alleviating the limitation that previous 3D pose estimators easily overfit in appearance and camera views, we develop a data augmentation algorithm to enrich the appearance model with in-the-wild scenarios and the geometry model with virtual 2D poses under unseen camera views. The proposed algorithm efficiently improves model robustness against appearance variations and cross-view generalization ability. Additionally, we propose a new experimental protocol on *Human3.6M* that follows a cross-view setting. This evaluation protocol focuses on the model robustness against camera view variance and has been long-time ignored by previous protocols. We conducted exhaustive experiments on public human pose benchmarks, including *Human3.6M*, *HumanEva*, *HHOI* and *MPII*, to verify the generalization issues of existing methods, and evaluate the proposed method for cross-view human pose estimation. Results show that our method can significantly reduce pose estimation errors and clearly outperform the alternative methods. We will explore more expressive and interpretable pose grammar representations, more effective and efficient network architectures and better data augmentation strategies in the future.

Acknowledgment

This work is supported by DARPA XAI Project N66001-17-2-4029, N66001-17-2-3602, ONR MURI Project N00014-16-1-2007, JUMP CRISP center, NSF IIS 1423305, 1657600 and ARO W911NF1810296. The authors would like to thank Bruce Xiaohan Nie, Ping Wei and Hao-Shu Fang for their prior efforts and supports.

References

[1] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 2014.

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.

[3] J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman, "Personalizing human video pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3063–3072.

[4] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*, 2016, pp. 483–499.

[5] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall, "A dual-source approach for 3d pose estimation from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4948–4956.

[6] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *IEEE International Conference on Computer Vision*, 2017, pp. 2659–2668.

[7] L. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International Journal of Computer Vision*, vol. 87, no. 1, pp. 4–27, 2010.

[8] T. Shu, M. S. Ryoo, and S. Zhu, "Learning social affordance for human-robot interaction," in *International Joint Conference on Artificial Intelligence*, 2016.

[9] B. X. Nie, P. Wei, and S.-C. Zhu, "Monocular 3d human pose estimation by predicting depth on joints," in *IEEE International Conference on Computer Vision*, 2017.

[10] H.-S. Fang, Y. Xu, W. Wang, and S.-C. Zhu, "Learning pose grammar to encode human body configuration for 3d pose estimation," in *AAAI Conference on Artificial Intelligence*, 2018.

[11] G. S. Paul, P. Viola, and T. Darrell, "Fast pose estimation with parameter-sensitive hashing," in *IEEE International Conference on Computer Vision*, 2003, pp. 750–757.

[12] H. Jiang, "3d human pose reconstruction using millions of exemplars," in *IEEE International Conference on Pattern Recognition*, 2010.

[13] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, "3d human pose estimation: A review of the literature and analysis of covariates," *Computer Vision and Image Understanding*, vol. 152, pp. 1–20, 2016.

[14] H. Rhodin, J. Sporri, I. Katircioglu, V. Constantin, F. Meyer, E. Muller, M. Salzmann, and P. Fua, "Learning monocular 3d human pose estimation from multi-view images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[15] M. Kocabas, S. Karagoz, and E. Akbas, "Self-supervised learning of 3d human pose using multi-view geometry," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[16] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov, "Learnable triangulation of human pose," in *IEEE International Conference on Computer Vision*, 2019.

[17] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis, "Sparseness meets deepness: 3d human pose estimation from monocular video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4966–4975.

[18] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[19] Y. Cheng, B. Yang, B. Wang, W. Yan, and R. T. Tan, "Occlusion-aware networks for 3d human pose estimation in video," in *IEEE International Conference on Computer Vision*, 2019.

[20] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[21] Y. Xu, S.-C. Zhu, and T. Tung, "Denserac: Joint 3d pose and shape estimation by dense render-and-compare," in *IEEE International Conference on Computer Vision*, 2019.

[22] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *IEEE International Conference on Computer Vision*, 2019.

[23] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, "3d human pose estimation in the wild by adversarial learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[24] D. C. Luvizon, D. Picard, and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[25] M. R. I. Hossain and J. J. Little, "Exploiting temporal information for 3d human pose estimation," in *European Conference on Computer Vision*, 2018.

[26] H. Rhodin, M. Salzmann, and P. Fua, "Unsupervised geometry-aware representation for 3d human pose estimation," in *European Conference on Computer Vision*, 2018.

[27] A. Nibali, Z. He, S. Morgan, and L. Prendergast, "3d human pose estimation with 2d marginal heatmaps," in *IEEE Winter Conference on Applications of Computer Vision*, 2019.

[28] S. Li and A. B. Chan, "3d human pose estimation from monocular images with deep convolutional neural network," in *Asian Conference on Computer Vision*, 2014, pp. 332–347.

[29] S. Li, W. Zhang, and A. B. Chan, "Maximum-margin structured learning with deep networks for 3d human pose estimation," in *IEEE International Conference on Computer Vision*, 2015, pp. 2848–2856.

[30] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua, "Structured prediction of 3d human pose with deep neural networks," in *British Machine Vision Conference*, 2016.

[31] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose,"

in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1263–1272.

[32] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *IEEE International Conference on Computer Vision*, 2017, pp. 2621–2630.

[33] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3d human pose estimation in the wild: a weakly-supervised approach," in *IEEE International Conference on Computer Vision*, 2017, pp. 398–407.

[34] M. Veges, V. Varga, and A. Lorincz, "3d human pose estimation with siamese equivariant embedding," *Neurocomputing*, vol. 339, pp. 194 – 201, 2019.

[35] C. Li and G. H. Lee, "Generating multiple hypotheses for 3d human pose estimation with mixture density network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[36] I. Habibie, W. Xu, D. Mehta, G. Pons-Moll, and C. Theobalt, "In the wild human pose estimation using explicit 2d features and intermediate 3d representations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[37] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3d human pose regression," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[38] B. Wandt and B. Rosenhahn, "Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[39] H. Gaifman, "Dependency systems and phrase-structure systems," *Information and control*, vol. 8, no. 3, pp. 304–337, 1965.

[40] T. Liu, S. Chaudhuri, V. Kim, Q. Huang, N. Mitra, and T. Funkhouser, "Creating consistent scene graphs using a probabilistic grammar," *ACM Transactions on Graphics*, vol. 33(6), pp. 1–12, 2014.

[41] Y. Xu, X. Liu, L. Qin, and S.-C. Zhu, "Cross-view people tracking by scene-centered spatio-temporal parsing," in *AAAI Conference on Artificial Intelligence*, 2017.

[42] S. Fidler, G. Berginc, and A. Leonardis, "Hierarchical statistical learning of generic parts of object structure," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 182–189.

[43] G. Gazdar, *Generalized phrase structure grammar*. Harvard University Press, 1985.

[44] O. Firschein, "Syntactic pattern recognition and applications," *Proceedings of the IEEE*, vol. 71, no. 10, pp. 1231–1231, 1983.

[45] S. Geman, D. F. Potter, and Z. Chi, "Composition systems," *Quarterly of Applied Mathematics*, vol. 60, no. 4, pp. 707–736, 2002.

[46] S.-C. Zhu, D. Mumford *et al.*, "A stochastic grammar of images," *Foundations and Trends in Computer Graphics and Vision*, vol. 2, no. 4, pp. 259–362, 2006.

[47] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[48] B. Rothrock, S. Park, and S.-C. Zhu, "Integrating grammar and segmentation for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3214–3221.

[49] S. Park, X. Nie, and S.-C. Zhu, "Attributed and-or grammar for joint parsing of human pose, parts and attributes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 7, pp. 1555–1569, 2018.

[50] W. Wang, Y. Xu, J. Shen, and S.-C. Zhu, "Attentive fashion grammar network for fashion landmark detection and clothing category classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[51] X. Liu, Y. Zhao, and S.-C. Zhu, "Single-view 3d scene reconstruction and parsing by attribute grammar," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 710–725, 2018.

[52] H. Qi, Y. Xu, T. Yuan, T. Wu, and S.-C. Zhu, "Scene-centric joint parsing of cross-view videos," in *AAAI Conference on Artificial Intelligence*, 2018.

[53] S. Huang, S. Qi, Y. Zhu, Y. Xiao, Y. Xu, and S.-C. Zhu, "Holistic 3d scene parsing and reconstruction from a single rgb image," in *European Conference on Computer Vision*, 2018.

[54] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1165–1179, 2017.

[55] Y. Xu, L. Qin, X. Liu, J. Xie, and S.-C. Zhu, "A causal and-or graph model for visibility fluent reasoning in tracking interacting objects," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[56] D. G. Hays, "Dependency theory: A formalism and some observations," *Language*, vol. 40, no. 4, pp. 511–525, 1964.

[57] S. Park, X. Nie, and S.-C. Zhu, "Attributed and-or grammar for joint parsing of human pose, parts and attributes," in *IEEE International Conference on Computer Vision*, 2015, pp. 1555–1569.

[58] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[59] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.

[60] A. Collins and J. Valentine, "Defining phyla: evolutionary pathways to metazoan body plans," *EVOLUTION & DEVELOPMENT*, vol. 3, no. 6, pp. 432–442, 2001.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[62] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial Intelligence and Statistics*, 2015.

[63] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *British Machine Vision Conference*, 2010.

[64] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[65] F. P. J. van der Hulst, S. Schätzle, C. Preusche, and A. Schiele, "A functional anatomy based kinematic human hand model with simple size adaptation," in *IEEE International Conference on Robotics and Automation*, 2012, pp. 5123–5129.

[66] K. Yamane and Y. Nakamura, "Robot kinematics and dynamics for modeling the human body," in *Robotics Research*, 2011, pp. 49–60.

[67] D. Lura, M. Wernke, S. Carey, R. Alqasemi, and R. Dubey, "Inverse kinematics of a bilateral robotic human upper body model based on motion analysis data," in *IEEE International Conference on Robotics and Automation*, 2013, pp. 5303–5308.

[68] X. Lan and D. P. Huttenlocher, "Beyond trees: common-factor models for 2d human pose recovery," in *IEEE International Conference on Computer Vision*, 2005, pp. 470–477.

[69] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient matching of pictorial structures," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2000.

[70] K. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Association for Computational Linguistics*, 2015.

[71] Y. Liu, H. Hel-Or, C. S. Kaplan, and L. Van Gool, "Computational symmetry in computer vision and computer graphics: A survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 5, no. 1–2, pp. 1–199, 2010.

[72] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and reidentification," *Computer Vision and Image Understanding*, vol. 117, no. 2, pp. 130–144, 2010.

[73] D. Eigen, C. Puhrsch, and R. Fergus, "Beyond planar symmetry: Modeling human perception of reflection and rotation symmetries in the wild," in *IEEE International Conference on Computer Vision*, 2017.

[74] T. Flash and B. Hochner, "Motor primitives in vertebrates and invertebrates," *Current Opinion in Neurobiology*, vol. 15, no. 6, pp. 660–666, 2005.

[75] T. Flash, Y. Meirovitch, and A. Barliya, "Defining phyla: evolutionary pathways to metazoan body plans," *Journal Robotics and Autonomous Systems*, vol. 61, no. 4, pp. 330–339, 2013.

[76] E. Chiovetto and M. A. Giese, "Kinematics of the coordi-nation of pointing during locomotion," *Plos One*, vol. 8, no. 11, 2013.

[77] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[78] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua, "Direct prediction of 3d body poses from motion compensated sequences," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 991–1000.

[79] Y. Du, Y. Wong, Y. Liu, F. Han, Y. Gui, Z. Wang, M. Kankanhalli, and W. Geng, "Marker-less 3d human motion capture with monocular image sequence and height-maps," in *European Conference on Computer Vision*, 2016, pp. 20–36.

[80] M. Sanzari, V. Ntouskos, and F. Pirri, "Bayesian image based 3d pose estimation," in *European Conference on Computer Vision*, 2016, pp. 566–582.

[81] C.-H. Chen and D. Ramanan, "3d human pose estimation = 2d pose estimation+ matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5759–5767.

[82] F. Moreno-Noguer, "3d human pose estimation from a single image via distance matrix regression," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1561–1570.

[83] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3d pose estimation from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5689–5698.

[84] B. Tekin, P. Marquez Neila, M. Salzmann, and P. Fua, "Learning to fuse 2d and 3d image cues for monocular body pose estimation," in *IEEE International Conference on Computer Vision*, 2017, pp. 3961–3970.

[85] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3d human pose reconstruction," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1446–1455.

[86] X. Zhou, M. Zhu, S. Leonardos, and K. Daniilidis, "Sparse representation for 3d shape estimation: A convex relaxation approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1648–1661, 2017.

[87] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *European Conference on Computer Vision*, 2016, pp. 561–578.

[88] I. Kostrikov and J. Gall, "Depth sweep regression forests for estimating 3d human pose from images," in *British Machine Vision Conference*, 2014.

[89] G. Rogez and C. Schmid, "Mocap-guided data augmentation for 3d pose estimation in the wild," in *Annual Conference on Neural Information Processing Systems*, 2016, pp. 3108–3116.

[90] C. Ionescu, F. Li, and C. Sminchisescu, "Latent structured models for human pose estimation," in *IEEE International Conference on Computer Vision*, 2011.

[91] V. Ramakrishna, T. Kanade, and Y. Sheikh, "Reconstructing 3d human pose from 2d image landmarks," in *European Conference on Computer Vision*, 2012, pp. 573–586.

[92] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer, "Single image 3d human pose estimation from noisy observations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2673–2680.

[93] G. W. Taylor, L. Sigal, D. J. Fleet, and G. E. Hinton, "Dynamical binary latent variable models for 3d human pose tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 631–638.

[94] L. Bo and C. Sminchisescu, "Twin gaussian processes for structured prediction," *International Journal of Computer Vision*, vol. 87, no. 1, pp. 28–52, 2010.

[95] L. Sigal, M. Isard, H. Haussecker, and M. J. Black, "Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation," *International Journal of Computer Vision*, vol. 98, no. 1, pp. 15–48, 2012.

[96] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer, "A joint model for 2d and 3d pose estimation from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3634–3641.

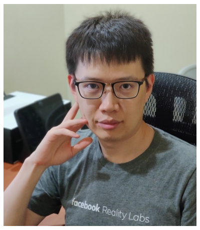
[97] I. Radwan, A. Dhall, and R. Goecke, "Monocular image 3d human pose estimation under self-occlusion," in *IEEE International Conference on Computer Vision*, 2013, pp. 1888–1895.

[98] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao, "Robust estimation of 3d human poses from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2361–2368.

[99] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic, "3d pictorial structures for multiple human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1669–1676.

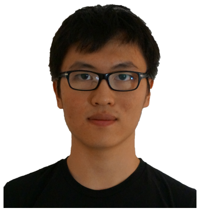
[100] A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt, "Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3810–3818.

[101] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.



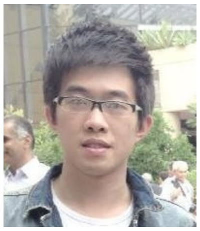
Yuanlu Xu is a research scientist at Facebook Reality Labs Research, Sausalito, USA. He received his Ph.D. degree from University of California, Los Angeles, USA, advised by Prof. Song-Chun Zhu. Before that, he received the master's degree from School of Information Science and Technology, Sun Yat-sen University, Guangzhou, China, advised by Prof. Liang Lin. He received the B.E. (Hons.) degree from School of Software, Sun Yat-sen University. His research interests include computer vision, 3D vision and graphics, machine learning, statistical modeling.

Wenguan Wang received his Ph.D. degree from Beijing Institute of Technology in 2018. He is currently a postdoc researcher at ETH Zurich, Switzerland. From 2016 to 2018, he was a visiting Ph.D. student in University of California, Los Angeles. From 2018 to 2019, he was a senior scientist at Inception Institute of Artificial Intelligence, UAE. His current research interests include computer vision, image processing and deep learning.



Tengyu Liu is a Ph.D. candidate at University of California at Los Angeles, USA. His current advisor is Prof. Song-Chun Zhu. Before that, he received the master's degree from University of California at Los Angeles, USA. He received the bachelor's degree from University of Illinois at Urbana-Champaign. His research interests are in computer vision and deep learning.

Xiaobai Liu is an Associate Professor of Computer Science at San Diego State University (SDSU), San Diego, USA. He received his PhD from the Huazhong University of Science and Technology (HUST), China. Before joining SDSU, he ever worked in National University of Singapore and University of California at Los Angeles. His research interests focus on the development of theories, algorithms, and models for the core computer vision problems. He has published 70 peer-reviewed articles in top-tier conferences (e.g. ICCV, CVPR, etc.) and leading journals (e.g.TPAMI, TIP etc.).



Jianwen Xie received his Ph.D. degree in statistics from University of California, Los Angeles (UCLA) in 2016. He is currently a senior research scientist at Baidu Research USA. Before joining Baidu, he was a senior research scientist at Hikvision Research Institute USA from 2017 to 2020, and a staff research associate and postdoctoral researcher in the Center for Vision, Cognition, Learning, and Autonomy (VCLA) at UCLA from 2016 to 2017. His research interests focus on generative modeling and learning with applications in computer vision.

Song-Chun Zhu received his Ph.D. degree from Harvard University in 1996, and worked at Brown, Stanford, Ohio State before joining UCLA in 2002 where he became professor of Statistics and Computer Science in 2006, and directed the UCLA Center for Vision, Cognition, Learning and Autonomy (VCLA). In 2020 he returned to China to establish the Beijing Institute for General Artificial Intelligence (BIGAI) as the founding director, and he was appointed Chair Professor at Tsinghua University and jointly Chair Professor and Dean of the AI Institute at Peking University. He has published over 300 papers in computer vision, statistical modeling and learning, cognitive science, natural language & situated dialogue, and robot autonomy, and commonsense reasoning. He received a number of honors, including the Helmholtz Test-of-time award in ICCV 2013, the Aggarwal prize from the Int'l Association of Pattern Recognition in 2008, the David Marr Prize in 2003 for image parsing with Z. Tu *et al.*, twice Marr Prize honorary nominations with Y. Wu *et al.*in 1999 for texture modeling and 2007 for object modeling respectively. He received the Sloan Fellowship in 2001, a US NSF Career Award in 2001, and an US ONR Young Investigator Award in 2001. He is a Fellow of IEEE since 2011, and served as the general co-chair for CVPR 2012 and CVPR 2019.