Learning Knowledge-guided Pose Grammar Machine for 3D Human Pose Estimation

Haoshu Fang[1,2*], **Yuanlu Xu**[1*], **Wenguan Wang**[1,3], **Xiaobai Liu**[4], **Song-Chun Zhu**[1]
[1]Dept. Computer Science and Statistics, University of California, Los Angeles (UCLA)
[2]Sch. Computer Science, Shanghai Jiao Tong University (SJTU)
[3]Sch. Computer Science, Beijing Institute of Technology (BIT)
[4]Dept. Computer Science, San Diego State University (SDSU)
fhaoshu@gmail.com, yuanluxu@cs.ucla.edu, wenguanwang@bit.edu.cn
xiaobai.liu@mail.sdsu.edu, sczhu@stat.ucla.edu

Abstract

In this paper, we propose a knowledge-guided pose grammar network to tackle the problem of 3D human pose estimation. Our model directly takes 2D poses as inputs and learns the generalized 2D-3D mapping function, which renders high applicability. The proposed network consists of a base network which efficiently captures pose-aligned features and a hierarchy of Bidirectional RNNs on top of it to explicitly incorporate a set of knowledge (e.g., kinematics, symmetry, coordination) and thus enforce high-level constraints over human poses. In learning, we develop a pose-guided sample simulator to augment training samples in virtual camera views, which further improves the generalization ability of our model. 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 ability of different methods. We empirically observe that most state-of-the-arts face difficulty under such setting while our method obtains superior performance.

1 Introduction

Estimating 3D human poses from a single still image has attracted attentions in the last few years because of its wide application potentials in robotics, autonomous vehicles, intelligent drones etc. This is a challenging reverse task since it aims to project 2D data back to 3D spaces and the inherent ambiguities are further enhanced with other factors, *e.g.*, clothes, occlusions, background clutters, etc. With the availability of large-scale pose datasets, *e.g.*, Human3.6M (Ionescu et al. 2014), deep networks based methods have achieved encouraging successes. These methods can be roughly divided into two categories. The first one is to learn end-to-end networks that project 2D input images to 3D poses directly, while the other category aims to extract 2D human poses from input images and then lift the 2D poses to 3D spaces.

There are many advantages to decouple the 3D human pose estimation problem into two stages. For the 2D pose estimation, for example, there exists a large amount of 2D pose data in the wild (Andriluka et al. 2014) and the 2D pose data can be easily labeled. The 2D pose estimation in the wild (Newell, Yang, and Deng 2016) is also mature enough to be deployed. For 2D to 3D projection, infinite 2D-3D pose pairs can be generated by projecting each 3D pose into 2D poses under different camera viewpoints and recent works (Yasin et al. 2016; Martinez et al. 2017) have shown that well-designed deep networks can achieve state-of-the-art result on Human3.6M dataset using only 2D pose data.

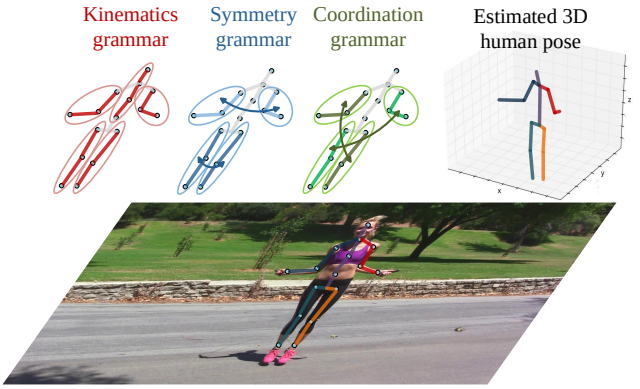

Figure 1: Illustration of human pose grammars, which express the knowledge of human embodiment, including kinematics, symmetry and motor coordination.

However, despite their promising results, few previous methods attached how to inject high-level human knowledge into current deep learning based detectors, which covers rich domain-specific information in this task. In this paper, we develop a deep grammar network to explicitly encode a set of knowledge over human poses, as illustrated in Figure 1. These knowledges explicitly express the composition process of joints-part-pose following various principles, including kinematics, symmetry and motor coordination, and serve as knowledge bases for constructing pose machine. In particular, we ground these knowledge in a multi-level RNN network which can be end-to-end trained with back-propagation.

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-arts heavily relies on the camera viewpoints. As shown in Table 1, once we change the split of training and testing set, using 3 cameras for training and testing on the forth camera, the performance of state-of-the-arts drops dramatically and is much worse than the image-based end-to-end method. These empirical studies suggested that existing methods might be over-fitting with the camera settings and bear poor generalization capabilities.

To tackle the above issues, in this work, we propose to study a cross-view 3D human estimation method, which aims to lift 2D poses to 3D spaces using deep networks on images of unseen camera views. More specifically, we develop a pose-guided simulator to augment training samples with virtual camera views, which can further improve system robustness. Our method is motivated by the previous works on learning by synthesis. Differently, we focus on the sampling of 2D pose instance from a given 3D space, following the basic geometry principles. In particular, we develop a pose-guided simulator to effectively generate training samples from unseen camera views. These samples can greatly reduce the risk of overfitting and thus improve generalization capabilities of the developed post estimation system.

We conducted exhaustive experiments on public human pose benchmarks, *e.g.*, Human3.6M, HumanEva, 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.

Contributions. There are two major contributions of the proposed framework: i) we propose a deep grammar network that incorporates both powerful encoding capabilities of deep neural networks and high-level semantic relations in human pose; ii) we propose a learning method that can improve generalization ability of current 2-step methods, allowing it to catch up with end-to-end methods or even outperforms image-based competitors.

2 Related Work

The proposed method is closely related to the following three research streams in computer vision and artificial intelligence.

3D pose estimation, basically, can be classified into two frameworks, *i.e.*, directly learn the 3D pose structure from 2D images, or break the problem into two cascaded tasks of first performing 2D pose estimation then reconstructing 3D pose from the estimated 2D joints. Specifically, for the first method, (Li and Chan 2014) proposed a multi-task convolutional network that simultaneously learns pose regression and part detection. (Tekin et al. 2016a) first learned an auto-encoder that describes 3D pose in high dimensional space then projected the input image to that space using CNN. (Pavlakos et al. 2017) represented 3D joints as points in a discretized 3D space and proposed a coarse-to-fine approach for iterative refinement. (Zhou et al. 2017) mixed 2D and 3D data and trained an unified network with two-stage cascaded structure. These methods heavily relies on well-labeled image and 3D ground-truth pairs, since they need to learn depth information from images.

To avoid this limitation, some work (Paul, Viola, and Darrell 2003; Jiang 2010; Yasin et al. 2016) tried to address this problem in a two step manner. For example, in (Yasin et al. 2016), the authors proposed an exemplar-based method to retrieve the nearest 3D pose in the 3D pose library using the estimated 2D pose. Recently, (Martinez et al. 2017) proposed a network that directly regresses 3D keypoints from 2D joint detections and achieves state-of-the-art performance. Our work takes a further step towards a unified 2D-to-3D projection network that integrates the learning power of deep learning and the domain-specific knowledge represented by hierarchy grammar model. The proposed method would offer a deep insight into the rationale behind this problem.

Grammar model receives heated endorsement due to its effectiveness in modeling diverse tasks (Liu, Mu, and Lin 2016; Liu et al. 2016; Xia et al. 2016; Park, Nie, and Zhu 2017). In (Han and Zhu 2009), the author approached the problem of image parsing using a stochastic grammar model. After that, generative grammar models are introduced in (Zhao and Zhu 2011) and (Liu et al. 2014) respectively for scene parsing. (Pero et al. 2013) further built a generative scene grammar to model the constitutionality of Manhattan structures in indoor scenes. In this paper, our representation can be analogized as a hierarchical attributed grammar model, with similar hierarchical structures, composition criteria as production rules, and soft constraints as probabilistic grammars. The difference lies in that our model is fully recursive and without semantics in middle levels.

3 Representation

We represent the 2D human pose E as a set of N_E joint locations

$$E = \{e_i : \ i = 1, \dots, N_E, \ e_i \in \mathbb{R}^2\}. \tag{1}$$

Our task is to estimate the corresponding 3D human pose $\mathbf{E}$ in the world reference frame. Suppose the 2D coordinate of a joint e_i is $[u_i, v_i]$ and the 3D coordinate $\mathbf{e}_i$ is $[X_i, Y_i, Z_i]$, we can describe the relation between 2D and 3D as a pinhole image projection

$$\begin{bmatrix} u_i \\ v_i \\ w_i \end{bmatrix} = K\,[R|RT] \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix}, K = \begin{bmatrix} \alpha_x & 0 & u_0 \\ 0 & \alpha_y & v_0 \\ 0 & 0 & 1 \end{bmatrix}, T = \begin{bmatrix} T_x \\ T_y \\ T_z \end{bmatrix}, \tag{2}$$

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 u_0 and v_0), R and T are camera extrinsic parameters of rotation and translation, respectively. Note we omit camera distortion for simplicity.

It involves 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→3D mapping $\mathbf{E} = f(E; \theta)$, where $f(\cdot)$ is a multi-to-multi mapping function, parameterized by θ.

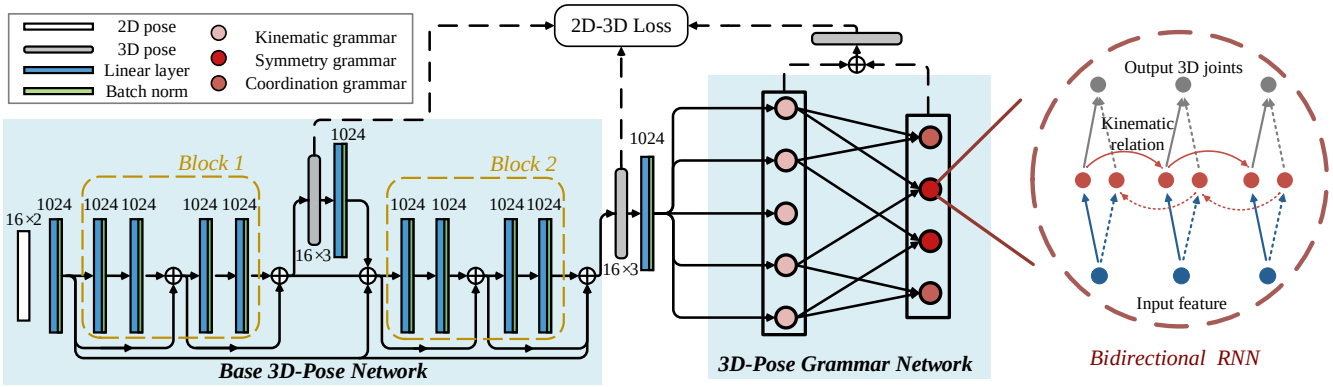

Figure 2: The proposed Pose Grammar Machine. Our model consists of two major components: a base network constituted by two basic blocks and a pose grammar network constituted by kinematics, symmetry and coordination grammars. Each grammar is represented as a Bidirectional RNN among certain joints. See text for detailed explanations.

3.1 Model Overview

Our model follows the line that directly estimating 3D human keypoints from 2D joint detections, which renders our model high applicability. More specifically, we extend various human pose grammars into deep neural network, where a basic 3D pose detection network is first used for extracting pose-aligned features, and a hierarchy of RNNs is built for encoding high-level 3D pose grammars for generating final reasonable 3D pose estimations. Above two networks work in a cascaded way, resulting in a strong 3D pose machine that inherits the representation power of neural network and high-level knowledge of articulated human body.

3.2 Base 3D-Pose Network

For building a solid foundation for high-level grammar model, we first use a base network for capturing well both 2D and 3D pose-aligned features. The base network is inspired by (Martinez et al. 2017), which has been demonstrated 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 with Batch Normalization, Dropout layers, and *ReLU* activation, are stacked for efficiently mapping the 2D-pose features to higher-dimensions. The input 2D pose detections E (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-dimension 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 take a full use of the information of initial 2D pose detections, we introduce *residual connections* (He et al. 2016) 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 direct access to the gradients from the loss function (detailed in Sec. 4), leading to an implicit deep supervision (Lee et al. 2015). 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 powerful representation that has well-aligned 3D-pose information and preserves the original 2D-pose information. Then we feed this feature into our 3D-pose grammar network.

3.3 3D-Pose Grammar Network

So far, our base network directly estimated the depth of each joint from the 2D pose detections. However, the natural of human body that rich inherent structures are involved in this task, motivates us to reason the 3D structure of the whole person in a global manner. Here we extend Bidirectional RNNs (Schuster and Paliwal 1997) (BRNN) to model high-level knowledge of 3D human pose grammar, which towards a more reasonable and powerful 3D pose machine 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. Basically, given a human body, we consider the following three types of grammars in our network.

Kinematic grammar $\mathcal{G}^{kin}$ describes human body movements without considering forces (*i.e.*, the red skeleton in Figure 1)). We define 5 kinematic grammars to represent the constraints among kinematically connected joints:

$$\begin{aligned}
\mathcal{G}^{kin}_{spine} &: \textit{head} \leftrightarrow \textit{thorax} \leftrightarrow \textit{spine} \leftrightarrow \textit{hip},\\
\mathcal{G}^{kin}_{l.arm} &: \textit{l.shoulder} \leftrightarrow \textit{l.elbow} \leftrightarrow \textit{l.wrist},\\
\mathcal{G}^{kin}_{r.arm} &: \textit{r.shoulder} \leftrightarrow \textit{r.elbow} \leftrightarrow \textit{r.wrist},\\
\mathcal{G}^{kin}_{l.leg} &: \textit{l.hip} \leftrightarrow \textit{l.knee} \leftrightarrow \textit{l.foot},\\
\mathcal{G}^{kin}_{r.leg} &: \textit{r.hip} \leftrightarrow \textit{r.knee} \leftrightarrow \textit{r.foot}.
\end{aligned} \tag{3}$$

Kinematic 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 $\mathcal{G}^{sym}$ measure bilateral symmetry of human body (*i.e.*, blue skeleton in Figure 1), as human body can be divided into matching halves by drawing a line down the center; the left and right sides are mirror images of each other.

$$\begin{aligned} \mathcal{G}^{sym}_{arm} &: \mathcal{G}^{kin}_{l.arm} \leftrightarrow \mathcal{G}^{kin}_{r.arm} , \\ \mathcal{G}^{sym}_{leg} &: \mathcal{G}^{kin}_{l.leg} \leftrightarrow \mathcal{G}^{kin}_{r.leg} . \end{aligned} \tag{4}$$

Coordination grammar $\mathcal{G}^{crd}$ represents movements of several limbs combined in a certain manner (*i.e.*, green skeleton in Figure 1). In this paper, we consider simplified motor coordination between human arm and leg. We define 2 coordination grammars to represent constraints on people coordinated movements:

$$\begin{aligned} \mathcal{G}^{crd}_{l.arm\&r.leg} &: \mathcal{G}^{kin}_{l.arm} \leftrightarrow \mathcal{G}^{kin}_{r.leg} , \\ \mathcal{G}^{crd}_{r.arm\&l.leg} &: \mathcal{G}^{kin}_{r.arm} \leftrightarrow \mathcal{G}^{kin}_{l.leg} . \end{aligned} \tag{5}$$

The RNN naturally supports chain-like structure, which provides a powerful tool for modeling our grammar formulations with deep learning. Classical BRNN is selected as our basic building block, due to the bidirectional dependencies of our grammars. There are two states (forward/backfward directions) encoded in BRNN. At each time step t, with the input feature x_t, the output y_t is determined by considering two-direction states h^f_i and h^b_i:

$$y_t = \phi(W^f_y h^f_t + W^b_y h^b_t + b_y), \tag{6}$$

where ϕ is the softmax function and the states h^f_t, h^b_t are computed as:

$$\begin{aligned} h^f_t &= \tanh(W^f_h h^f_{t-1} + W^f_x x_t + b^f_h) , \\ h^b_t &= \tanh(W^b_h h^b_{t+1} + W^b_x x_t + b^b_h) . \end{aligned} \tag{7}$$

As shown in Figure 2, we build a two-layer tree-like hierarchy of BRNNs for modeling our three grammars, where each of the BRNNs shares same equation in Equ. 6 and the three grammars are represented by the edges between BRNNs nodes or implicitly encoded in to BRNN architecture.

For the bottom layer, five BRNNs are built for modeling the five relations defined in kinematics grammar (Equ. 3). 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/backfoward propagated efficiently over the two states with BRNN, thus the five Kinematics relations are implicitly modeled by the bidirectional 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, totally four BRNN nodes are derived, two for symmetry relations (Equ. 4) and two for coordination dependencies (Equ. 4). 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 input, and produces estimations for the six 3D-joints taking their symmetry relations into account. Similarly, for the coordination nodes, such as $\mathcal{G}^{crd}_{l.arm\&r.leg}$, it leverages the estimations from $\mathcal{G}^{kin}_{l.arm}$ and $\mathcal{G}^{kin}_{r.leg}$ BRNNs and refines the 3D joints estimations according to coordination grammar.

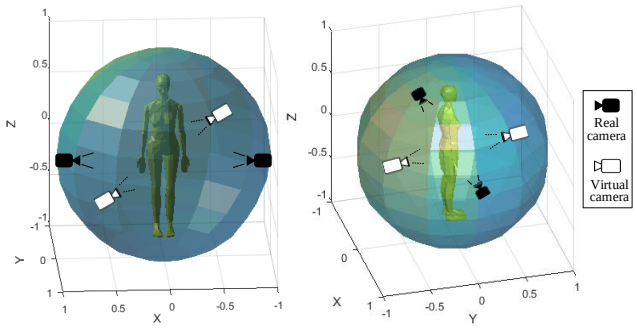

Figure 3: Illustration of virtual camera simulation. The black camera icons stand for real camera settings while the white camera icons simulated virtual camera settings.

In this way, we inject three kinds of human pose grammar into a tree-BRNN model and the final 3D human joints estimations are achieved by mean-pooling the results from all the nodes in the grammar hierarchy.

4 Learning

Given a training set Ω:

$$\Omega = \{(\hat{E}_i, \hat{\mathbf{E}}_i) : i = 1, \ldots, N_\Omega\}, \tag{8}$$

where $\hat{E}_i$ and $\hat{\mathbf{E}}_i$ denote ground-truth 2D and 3D pose pairs, we define the 2D-3D loss of learning the mapping function $f(E;\theta)$ as

$$\begin{aligned} \theta^* &= \arg\min_\theta \ell(\Omega|\theta) \\ &= \arg\min_\theta \sum_{i=1}^{N_\Omega} \|f(\hat{E}_i|\theta) - \hat{\mathbf{E}}_i\|_2. \end{aligned} \tag{9}$$

The loss measures the Euclidian distance between predicted 3D pose and true 3D pose.

The entire learning process consists of two steps: i) learning basic blocks in the base network with 2D-3D loss. ii) attaching pose grammar network on top of the trained base network, and fine-tune the whole network in an end-to-end manner.

4.1 Pose-guided Sample Simulator

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 (4 on average) recording the human subject. This raises the doubt whether learning on such dataset can lead to a generalized

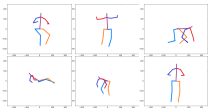

Figure 4: Examples of learned 2D atomic poses in probability distribution $p(E|\hat{E})$.

3D pose estimator applicable in other scenes with different camera positions. We believe that a data augmentation process will help improve the model performance and generalization ability. Therefore, we propose a novel Pose-guided Sample Simulator (PGSS) to generate additional training samples. The generation process consist of two steps: i) projecting ground-truth 3D pose $\hat{\mathbf{E}}$ onto virtual camera planes to obtain ground-truth 2D pose $\hat{E}$, ii) simulating 2D pose detections E by sampling conditional probability distribution $p(E|\hat{E})$.

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 $\hat{\mathbf{E}}$ onto the virtual camera plane and obtain the corresponding ground-truth 2D pose $\hat{E}$.

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

$$p(E|\hat{E}) = p(\epsilon) = \sum_{i=1}^{N_G} \omega_i \, \mathbb{N}(\epsilon; \mu_i, \Sigma_i), \tag{10}$$

where $\epsilon = E - \hat{E}$, N_G denotes the number of Gaussian distributions, ω_i denotes a combination weight for the i-th component, $\mathbb{N}(\epsilon; \mu_i, \Sigma_i)$ denotes the i-th multivariate Gaussian distribution with mean μ_i and covariance Σ_i. As suggested in (Andriluka et al. 2014), we set $N_G = 42$. For efficiency issues, the covariance matrix Σ_i is assumed to be in the form:

$$\Sigma_i = \begin{bmatrix} \sigma_{i,1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_{i,j} \end{bmatrix}, \quad \sigma_{i,j} \in \mathbb{R}^{2\times 2} \tag{11}$$

where $\sigma_{i,j}$ is the covariance matrix for joint $e_{i,j}$. This constraint enforces independence among each joint $e_{i,j}$ in 2D pose E_i.

The probability distribution $p(E|\hat{E})$ 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_i 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 $\hat{E}$, we sample $p(E|\hat{E})$ to generate simulated detections E and thus use it augment the training set Ω. By doing so we mitigate the discrepancy between the training data and the testing data. The effectiveness of our proposed PGSS is validated in Section 5.5.

5 Experiments

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

5.1 Datasets

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

Human3.6M (Ionescu et al. 2014) 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.*, greeting, eating and walking). The 3D pose ground-truth is captured by a motion capture (Mocap) system and all camera parameters (intrinsic and extrinsic parameters) are provided.

HumanEva-I (Sigal, Balan, and Black 2010) is another widely used dataset for human 3D pose estimation, which is also collected in a controlled indoor environment using a Mocap system. *HumanEva-I* dataset has fewer subjects and actions, compared with *Human3.6M* dataset.

MPII (Andriluka et al. 2014) 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.

5.2 Evaluation Protocols

For **Human3.6M**, the standard protocol is using all 4 camera views in subjects S1, S5, S6 and S7 for training and the same 4 camera views in subjects S9 and S11 for testing. This standard protocol is called *protocol #1*. In some works, the predictions are post-processed via a rigid transformation before comparing to the ground-truth, which is referred as *protocol #2*.

In above two protocols, the same 4 camera views are both used for training and testing. This raise the question whether or not 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. In our setting, subjects S1, S5, S6 and S7 in 3 camera views are used for training while subjects S9 and S11 in the other camera view are selected for testing. The suggested protocol guarantees that not only subjects but also camera views are different for

Protocol #1	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
LinKDE (PAMI'16)	132.7	183.6	132.3	164.4	162.1	205.9	150.6	171.3	151.6	243.0	162.1	170.7	177.1	96.6	127.9	162.1
Tekin et al. (ICCV'16)	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
Du et al. (ECCV'16)	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
Chen & Ramanan (Arxiv'16)	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)	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
Zhou et al. (ICCV'17)	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)	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
Ours	**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**
Protocol #2	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Ramakrishna et al.(ECCV'12)	137.4	149.3	141.6	154.3	157.7	158.9	141.8	158.1	168.6	175.6	160.4	161.7	150.0	174.8	150.2	157.3
Bogo et al.(ECCV'16)	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
Moreno-Noguer (CVPR'17)	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)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	51.9
Martinez et al. (ICCV'17)	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
Ours	**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**
Protocol #3	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Pavlakos et al. (CVPR'17)	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)	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)	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	**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**

Table 1: Quantitative comparisons of Average Euclidean Distance (mm) between the estimated pose and the ground-truth on *Human3.6M* under *Protocol #1*, *Protocol #2* and *Protocol #3*. The best score is marked in **bold**.

training and testing, eliminating interferences of subject appearance and camera parameters, respectively. We refer our new protocol as *protocol #3*.

For **HumanEva-I**, we follow the previous protocol, evaluating on each action separately with all subjects. A rigid transformation is performed before computing the mean reconstruction error.

5.3 Implementation Details

We implement our method using Keras with Tensorflow as backend. We first train our base network for 200 epoch. The learning rate is set as 0.001 with exponential decay and the batch size is set to 64 in the first step. Then we add the 3D-Pose Grammar Network on top of the base network and fine-tune the whole network together. The learning rate is set as 10^{-5} during the second step to guarantee model stability in the training phase. We adopt Adam (Kingma and Ba 2014) for optimization for both steps.

We perform 2D pose detections using a state-of-the-art 2D pose estimator (Newell, Yang, and Deng 2016). We fine-tuned the model on *Human3.6M* and employ the pre-trained model on *HumanEva-I* and *MPII*. Our Pose Grammar Machine is trained with 2D pose detections as inputs and 3D pose ground-truth as outputs. For *protocol #1* and *protocol #2*, the data augmentation is omitted due to little improvement and tripled training time. For *protocol #3*, 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 PCSS. 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 25 ms on a Titan X GPU and a forward pass takes 10 ms only, allowing us to train and test our network efficiently.

Methods	Walking			Jogging			Avg.
	S1	S2	S3	S1	S2	S3	
Simo-Serra et al. (CVPR'13)	65.1	48.6	73.5	74.2	46.6	32.2	56.7
Kostrikov et al. (BMVC'14)	44.0	30.9	41.7	57.2	35.0	33.3	40.3
Yasin et al. (CVPR'16)	35.8	32.4	41.6	46.6	41.4	35.4	38.9
Moreno-Noguer (CVPR'17)	19.7	**13.0**	**24.9**	39.7	20.0	21.0	26.9
Pavlakos et al. (CVPR'17)	22.3	19.5	29.7	28.9	21.9	23.8	24.3
Martinez et al. (ICCV'17)	19.7	17.4	46.8	**26.9**	18.2	18.6	24.6
Ours	**19.4**	16.8	37.4	30.4	**17.6**	**16.3**	**22.9**

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

5.4 Results and Comparisons

Human3.6M. We evaluate our method under all three protocols. We compare our method with 10 state-of-the-arts (Ionescu et al. 2014; Tekin et al. 2016b; Du et al. 2016; Chen and Ramanan 2016; Sanzari, Ntouskos, and Pirri 2016; Rogez and Schmid 2016; Bogo et al. 2016; Pavlakos et al. 2017; Zhou et al. 2017; Martinez et al. 2017) and report quantitative comparisons in Table 1. From the results, our method obtains superior performance over the competing methods under all protocols.

To verify our claims, we re-train three previous methods, which obtain top performance under *protocol #1*, with *protocol #3*. The quantitative results are reported in Table. 1. The large drop of performance (17% – 41%) of previous 2D-3D projection models: (Pavlakos et al. 2017; Zhou et al. 2017; Martinez et al. 2017), which demonstrates the blindspot of

Figure 5: Quantitative results of our method on *Human3.6M* and *MPII*. We show the estimated 2D pose on the original image and the estimated 3D pose from a novel view. Results on *Human3.6M* are drawn in the first row and results on *MPII* are drawn in the second to fourth row. Best viewed in color.

previous evaluation protocols and the over-fitting problem of those models.

Notably, our method greatly surpasses previous methods (-12 mm improvement over the second best under cross-view evaluation (*i.e.*, *protocol #3*). Additionally, the large performance gap of (Martinez et al. 2017) under *protocol #1* and *protocol #3* (62.9 *vs* 84.9) demonstrates that previous 2D-to-3D projection networks easily over-fit to camera views. Our general improvements over different settings demonstrate our superior performance and good generalization.

HumanEva-I. We compare our method with 6 state-of-the-arts (Simo-Serra et al. 2013; Kostrikov and Gall 2014; Yasin et al. 2016; Moreno-Noguer 2017; Pavlakos et al. 2017; Martinez et al. 2017). The quantitative comparisons on *HumanEva-I* are reported in Table 2. As seen, our results outperforms previous methods across the vast majority of subjects and on average.

MPII. 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.

5.5 Ablation studies

We study different components of our model on *Human 3.6M* dataset under *protocol #3*, as reported in Table 3.

Pose grammar. We first study the effectiveness of our grammar model, which encodes high-level grammar constraints into our network. First, we exam the performance of our baseline by removing all three grammars from our model, the error is $75.1mm$. Adding the kinematics grammar provides parent-child relations to body joints, reducing the error by 1.6% ($75.1mm \rightarrow 73.9mm$). Adding on top the symmetry grammar can obtain an extra error drops($73.9mm \rightarrow 73.2mm$). After combing all three grammars together, we can reach an final error of $72.8mm$.

Pose-guided Sample Simulator (PGSS). Next we evaluate the influence of our 2D-pose samples simulator. 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 an 7% errors drop ($82.6mm \rightarrow 76.7mm$),

Component	Variants	Error (mm)	Δ
	Ours, full	72.8	–
Pose grammars	w/o grammar	75.1	2.3
	w kinematics	73.9	1.1
	w kinematics+symmetry	73.2	0.4
PGSS	w/o extra 2D-3D pairs	82.6	9.8
	w extra 2D-3D pairs, GT	76.7	3.9
	w extra 2D-3D pairs, simple	78.0	5.2

Table 3: Ablation studies on different components in our method. The evaluation is performed on *Human3.6M* under *Protocol #3*. See text for detailed explanations.

showing that extra training data indeed expand the generalization ability. Next, we compare our Pose-guided Sample Simulator to a simple baseline, that is, generating samples by adding random noises to each joint, say an arbitrary Gaussian distribution or a white noise. Unsurprisingly, we observe a drop of performance, which is even 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. Therefore, this ablative study validates the effectiveness of our PGSS.

6 Conclusion

In this paper, we propose a pose grammar machine to encode the mapping function of human pose from 2D to 3D. By proposing a deeper base network and explicitly expressing human pose grammars, our method obtains superior performance over other state-of-the-arts. We will explore more effective and efficient network architectures in the future.

References

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

[Bogo et al. 2016] Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; and Black, M. J. 2016. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*.

[Chen and Ramanan 2016] Chen, C.-H., and Ramanan, D. 2016. 3d human pose estimation= 2d pose estimation+ matching. *arXiv preprint arXiv:1612.06524*.

[Du et al. 2016] Du, Y.; Wong, Y.; Liu, Y.; Han, F.; Gui, Y.; Wang, Z.; Kankanhalli, M.; and Geng, W. 2016. Marker-less 3d human motion capture with monocular image sequence and height-maps. In *European Conference on Computer Vision*.

[Han and Zhu 2009] Han, F., and Zhu, S. 2009. Bottom-up/top-down image parsing with attribute grammar. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(1):59–73.

[He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[Ionescu et al. 2014] Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2014. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(7):1325–1339.

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

[Kingma and Ba 2014] Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

[Lee et al. 2015] Lee, C.-Y.; Xie, S.; Gallagher, P.; Zhang, Z.; and Tu, Z. 2015. Deeply-supervised nets. In *Artificial Intelligence and Statistics*.

[Li and Chan 2014] Li, S., and Chan, A. B. 2014. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*.

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

[Liu et al. 2016] Liu, J.; Li, Y.; Allen, P.; and Belhumeur, P. 2016. Articulated pose estimation using hierarchical exemplar-based models.

[Liu, Mu, and Lin 2016] Liu, X.; Mu, Y.; and Lin, L. 2016. A stochastic grammar for fine-grained 3d scene reconstruction. In *International Joint Conference on Artificial Intelligence*.

[Martinez et al. 2017] Martinez, J.; Hossain, R.; Romero, J.; and Little, J. J. 2017. A simple yet effective baseline for 3d human pose estimation. In *IEEE International Conference on Computer Vision*.

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

[Newell, Yang, and Deng 2016] Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*.

[Park, Nie, and Zhu 2017] Park, S.; Nie, X.; and Zhu, S.-C. 2017. Attributed and-or grammar for joint parsing of human pose, parts and attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Paul, Viola, and Darrell 2003] Paul, G. S.; Viola, P.; and Darrell, T. 2003. Fast pose estimation with parameter-sensitive hashing. In *IEEE International Conference on Computer Vision*.

[Pavlakos et al. 2017] Pavlakos, G.; Zhou, X.; Derpanis, K. G.; and Daniilidis, K. 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *IEEE International Conference on Computer Vision*.

[Pero et al. 2013] Pero, L.; Bowdish, J.; Hartley, E.; Kermgard, B.; and Barnard, K. 2013. Understanding bayesian rooms using composite 3d object models. In *IEEE Conference on Computer Vision and Pattern Recognition*.

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

[Sanzari, Ntouskos, and Pirri 2016] Sanzari, M.; Ntouskos, V.; and Pirri, F. 2016. Bayesian image based 3d pose estimation. In *European Conference on Computer Vision*.

[Schuster and Paliwal 1997] Schuster, M., and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

[Sigal, Balan, and Black 2010] Sigal, L.; Balan, A. O.; and Black, M. J. 2010. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision* 87(1):4–27.

[Simo-Serra et al. 2013] Simo-Serra, E.; Quattoni, A.; Torras, C.; and Moreno-Noguer, F. 2013. A joint model for 2d and 3d pose estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[Tekin et al. 2016a] Tekin, B.; Katircioglu, I.; Salzmann, M.; Lepetit, V.; and Fua, P. 2016a. Structured prediction of 3d human pose with deep neural networks. *arXiv preprint arXiv:1605.05180*.

[Tekin et al. 2016b] Tekin, B.; Rozantsev, A.; Lepetit, V.; and Fua, P. 2016b. Direct prediction of 3d body poses from motion compensated sequences. In *IEEE International Conference on Computer Vision*.

[Xia et al. 2016] Xia, F.; Zhu, J.; Wang, P.; and Yuille, A. L. 2016. Pose-guided human parsing by an and/or graph using pose-context features. In *AAAI Conference on Artificial Intelligence*.

[Yasin et al. 2016] Yasin, H.; Iqbal, U.; Kruger, B.; Weber, A.; and Gall, J. 2016. A dual-source approach for 3d pose estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[Zhao and Zhu 2011] Zhao, Y., and Zhu, S. 2011. Image parsing via stochastic scene grammar. In *Annual Conference on Neural Information Processing Systems*.

[Zhou et al. 2017] Zhou, X.; Huang, Q.; Sun, X.; Xue, X.; and Wei, Y. 2017. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *IEEE International Conference on Computer Vision*.