Motion Projection Consistency-Based 3-D Human Pose Estimation With Virtual Bones From Monocular Videos

Guangming Wang, Graduate Student Member, IEEE, Honghao Zeng, Ziliang Wang, Zhe Liu, and Hesheng Wang, Senior Member, IEEE

Abstract—Real-time 3-D human pose estimation is crucial for human–computer interaction. It is cheap and practical to estimate 3-D human pose only from monocular video. However, the recent bone-splicing-based 3-D human pose estimation method brings about the problem of cumulative error. In this article, the concept of virtual bones is proposed to solve such a challenge. The virtual bones are imaginary bones between nonadjacent joints. They do not exist in reality, but they bring new loop constraints for the estimation of 3-D human joints. The proposed network in this article predicts real bones and virtual bones, simultaneously. The final length of real bones is constrained and learned by the loop constructed by the predicted real bones and virtual bones. Besides, the motion constraints of joints in consecutive frames are considered. The consistency between the 2-D projected position displacement predicted by the network and the captured real 2-D displacement by the camera is proposed as a new projection consistency loss for the learning of the 3-D human pose. The experiments on the Human3.6M data set demonstrate the good performance of the proposed method. Ablation studies demonstrate the effectiveness of the proposed interframe projection consistency constraints and intraframe loop constraints.

Index Terms—3-D human pose estimation, deep learning, motion constraints, virtual bones.

I. INTRODUCTION

In recent years, increasing attention is attracted to 3-D human pose estimation in videos, due to its wide application in the field of action recognition [1], robot learning [2]–[4], and robot control [5]–[8]. The state-of-the-art approaches [9]–[11] are mostly in two steps: 1) 2-D joint detection and 2) 3-D pose estimation from the 2-D joints.

The core of these methods is 3-D pose estimation based on 2-D joints, whose recognized difficulty is depth ambiguity. The information input is limited to the 2-D plane while the goal is to predict the joint coordinates in 3-D space, so the depth information needs to be compensated with other information, which nowadays is time and spatial information. Some approaches [9], [11]–[13] utilize the information of the adjacent frames and Chen et al. [10] utilized the information of the multiview. Specifically, Pavllo et al. [9] proposed an efficient approach for 3-D human pose estimation in video based on dilated temporal convolutions on 2-D keypoint trajectories. Chen et al. [10] first used the image-skeleton mapping module to obtain 2-D skeleton maps from images, and then a view synthesis module is used to predict 3-D pose. Lin and Lee [11] proposed a deep-learning-based framework that utilizes matrix factorization for sequential 3-D human pose estimation with the input 2-D joint position. Luvizon et al. [12] proposed a multitask framework that can estimate the 2-D and 3-D pose from images and recognize the action from the video sequence. Lee et al. [13] proposed a new long short-term memory (LSTM)-based deep learning architecture named propagating LSTM (p-LSTM) networks to infer depth. Compared with the method that directly predicts 3-D human pose from images, this two-stage work add the intermediate variables in the process of the prediction, which means introduce more constraints. However, joints still have a large degree of freedom, which is most pronounced at the end of the human body. Chen et al. [14] used consecutive frames to estimate the middle frame’s 3-D pose, which is decomposed into the length and direction prediction of bones. However, the method of obtaining joints by the accumulation of bones will accumulate errors.

To alleviate the accumulated errors, the first contribution in this article is to improve the bone prediction network by adding the prediction of virtual bones between nonadjacent joints to the bone prediction network. The loop constraint is constructed by real bones and virtual bones to reduce cumulative error and increase the prediction accuracy. Specifically, to avoid the overfitting caused by the limited number of actors in training data sets, each sampled frame is used to predict the corresponding 3-D joint position. The bone lengths are calculated from the estimated 3-D joint positions. To obtain the
real bone lengths in the current frame, a self-attention module is incorporated to weigh the real bone lengths from sampled frames. The virtual bone lengths are calculated from the 3-D joint positions in the current frame directly. The ground truth of bone length is used to optimize the self-attention module. The temporal convolutional network in [9] is used to predict the direction of all bones. The final positions of joints are derived from the bone length and direction of all bones through a fully connected network. The motivation to add the bones between nonadjacent joints is based on the idea that adding a proper amount of input virtual bones can increase the accuracy of the final joint prediction and reduce the overfitting to a certain extent.

The other contribution is proposing a new projection consistency loss. According to [15], even if the $l_1$ mean distance between the ground truth and estimated positions is the same, there will be a different distribution for joint positions in the time dimension. Some researchers [15], [16] make kinematics analysis and propose motion loss. Their ideas are all based on the continuity of the displacement of the joint, which is constrained by the real displacement in 3-D. A new projection consistency loss is proposed in this article, comparing the 2-D projection of the 3-D displacement of the estimated joints between adjacent frames and the 2-D displacement derived from the input 2-D keypoints. This loss can reduce the fluctuation of joint position estimation results in the continuous video, with no need of 3-D ground truth of joints.

In summary, the approach proposed in this article makes the following contributions.

1) A bone length prediction network with additional bones among nonadjacent joints is presented to avoid overfitting and predict joints more accurately.
2) A new projection loss based on the 2-D displacement of 3-D joints is proposed, not only smoothing the error between adjacent frames but also improving the accuracy of 3-D joint position prediction.
3) A variety of virtual bone combination modes is validated. Ablation studies demonstrate the effectiveness of the proposed method. The combination of virtual bones and projection loss lead to a good performance on the Human3.6M data set [17].

II. RELATED WORK

The development of the field of 3-D human pose estimation has undergone a variety of method changes. One of them uses neural networks to estimate the 3-D pose of the human body directly from the input image, including [18]–[20]. Mitra et al. [18] first trained the neural network in an unsupervised fashion with multiview data to get the human pose representation. Then, ground truth is used to map this representation to the 3-D human joints. It belongs to a semi-supervised method. Gabeur et al. [19] used the structure of encoder-decoder to estimate the depth of the front and rear surfaces of the human body as the representation of the 3-D pose. Considering the time information, Kocabas et al. [20] used the video as the input, obtaining 82 parameters of the skinned multiperson linear (SMPL) model [21] to represent the shape and pose of the human body.

Recently, the task of 3-D human joint estimation is divided into two parts. 2-D joint positions are first obtained from images, and then 3-D joint positions are predicted from the 2-D position. Methods for detecting 2-D joint positions from images have become mature [22], [23]. Zhang et al. [22] optimized the probability distribution obtained from the heat map, which leads to better 2-D joint estimation results. Nie et al. [23] proposed a chain method to represent joint coordinates. Each joint coordinate is represented by the root joint coordinate and the bone vector between adjacent joints.

With the help of mature technology for estimating 2-D keypoints, researchers estimate 3-D poses on this basis, such as [9], [13], [14], and [24]–[30]. Lee et al. [13] first used convolutional neural network to extract a 2-D pose from RGB image and then they proposed the p-LSTMs model to infer depth to obtain 3-D human pose. Wang et al. [24] noticed the importance of the occlusion relationship of the joint. Using the p-LSTMs model to infer the consistency of the poses from different viewing angles. Different from other works, Li et al. [26] estimated the length and direction angle of each bone and proposed a data set augmentation method that improves the accuracy of the algorithm in unusual poses.

Recent studies [9], [27]–[30] take into account the time information and use video as input instead of a separate frame. Pavllo et al. [9] regarded the process of estimating 3-D position from 2-D as the encoding part. The process of projecting 3-D back to 2-D is regarded as the decoding part. They train with the labeled data and use unlabeled data to calculate the consistency of input and decoded output as an unsupervised loss. Liu et al. [27] introduced an attention mechanism. Different frames are weighed and different convolution kernels are used to improve network structure and algorithm performance. Cheng et al. [28] applied the method of multiscale analysis in the time and space dimensions to deal with the problems of different sizes and different speeds of humans. Chen et al. [14] decomposed the 3-D pose into the bone length and the bone direction. Considering the length invariance of bones and the visibility of the joints on the image, they use the whole video combined with the attention mechanism to obtain a more accurate bone length estimation. At the same time, a new layered bone direction prediction network is proposed to get better results. Wu and Xiao [29] established a depth map of joints to calculate the limb depth map. Then, the hidden information extracted from the picture is combined to directly obtain 3-D poses. Jiang et al. [30] pretrained a series of 3-D poses using input pictures labeled with 2-D joint information. They take the weights of these poses from the model to get a rough pose and use the residual compensation to obtain the final predicted pose. To increase the accuracy, some researchers use the motion of the human joints in the video to improve the method of supervision.
For example, Xu et al. [16] took the kinematics analysis for monocular 3-D human pose estimation between multiple frames to correct the limbs at the end of the human body. Meanwhile, some researchers use a distance matrix to measure the relationship between different joints, such as [31]–[33]. Moreno-Noguer [31] proposed the Euclidean Distance Matrice (EDM) between 2-D keypoints to estimate the EDM between nonadjacent joints before estimating 3-D EDM. Gao et al. [32] based on [31] recovered the occluded joints before estimating 3-D EDM. Gao et al. [33] noticed the distance relationship between nonadjacent joints as [31]–[33]. Moreno-Noguer [31] proposed the Euclidean measure the relationship between different joints, such as left-elbow to neck, pelvis to right-elbow, and left-knee to right-knee.

Fig. 1. Overview of the proposed framework. It uses consecutive frames to predict the direction of real and virtual bones, randomly sampled frames to predict the length of real bones, and current frames to predict the length of virtual bones. The 3-D joint position is predicted by a fully connected network with predicted bones. Meanwhile, a projection consistency loss is used to constrain the learning of bone prediction.

Fig. 2. (a) Common skeleton of human with joints and bones for representation. (b) Schematic of virtual bones, such as left-elbow to neck, pelvis to right-elbow, and left-knee to right-knee.

III. MOTION PROJECTION CONSISTENCY-BASED 3-D HUMAN POSE ESTIMATION WITH VIRTUAL BONES

The overview of the proposed method is illustrated in Fig. 1. In this section, we elaborate on the details of our method. In Section III-A, we introduce the architecture of the bone prediction network with the virtual bone output. In Section III-B, we introduce the detail of obtaining the 3-D joint position from the bones. In Section III-C, we present the detail of projection consistency loss. In Section III-D, the other used losses are introduced.

A. Bone Length and Direction Prediction Networks

In recent years, the human skeleton structure commonly used by researchers is shown in Fig. 2(a), in which the bones between adjacent joints are recorded as real bones. In this article, we use the concept of virtual bones to represent the imaginary bones between nonadjacent joints, as in Fig. 2(b). These two terms will be used frequently in the following sections. In this article, the learning of real bones is optimized through the prediction of virtual bones and real bones simultaneously.

The structure of the bone length prediction network is shown in Fig. 3. The bone length prediction network not only predicts the bone length of real bones but also predicts the bone length of virtual bones. To get global information more effectively, like [14], we input 2-D joint location of J joints from f frames sampled from a video to the network. The 2-D joint locations are first used to predict coarse 3-D locations of the J joints, which are utilized to calculate the length of bones.

Because of the invariance of the bone length of real bones, the length of real bones is predicted for each of the random frames and weighed by an attention mechanism to get the length of real bones in the current frame

$$L_{\text{real}} = \sum_{f=1}^{f} w_i \psi_{\text{real}}(\phi(x_i))$$  (1)

where $w_i$ represents the matrix composed of the weight of each bone in the $i$th frame. $x_i$ ($i = 1, 2, \ldots, f$) represents the input 2-D joint location in random frames. $\psi_{\text{real}}$ represents the calculation to obtain the length of the real bones from coarse 3-D joint location in frames. $\phi$ represents the network that predicts the coarse 3-D joint location from random frames.

However, the bone length of virtual bone varies from frame to frame, so only the current frame is used, skipping the attention mechanism, to predict the bone length of $V$ virtual bones

$$L_{\text{virtual}} = \psi_{\text{virtual}}(\phi(x_{\text{current}}))$$  (2)

where $x_{\text{current}}$ represents the 2-D joint location in frame at the current timestamp. $\psi_{\text{virtual}}$ represents the calculation to obtain the length of virtual bones from coarse 3-D joint location in current frame. $\phi$ represents the network that predicts coarse 3-D joint location from the current frame.

The method for predicting bone direction is the same as [9], utilizing the temporal fully convolutional network whose output is the unit vector of bone direction

$$D_o = \psi(x_1, x_2, \ldots, x_f)$$  (3)

where $x_k$ ($k = 1, 2, \ldots, f$) represents the input 2-D joint location in consecutive frames. $\psi$ represents the temporal fully convolutional network.

For the standard 17-joint human skeleton structure, there is only one path from the root joint to the target joint. What our network predicts are the bone length and unit orientation of the bone, so with the bone vectors on the path from the root joint to the target joint, we only get the unique target joint coordinates. After the introduction of virtual bones, for a target joint, we will get multiple joint coordinates along different paths. The final target joint coordinate is obtained by weighting multiple predicted values, which can reduce the accumulate error and increase the prediction accuracy of the target joint.
Fig. 3. Detailed structure of bone length prediction network. The input of this network is the 2-D joint location $x$ in random $f$ frames from a video. $F^0$ is intermediate features. $n$ is the number of residual blocks. $F^1_k$ and $F^2_k$ are intermediate features in the $k$th residual block. $x_{current}$ is the 2-D joint location in the current frame. The 3-D joint location of all frames $\phi(x)$ and 3-D joint locations of current frame $\phi(x_{current})$ are obtained through the coarse joint location prediction network $\phi$. Then, real bone length $\phi_{real}(\phi(x))$ and virtual bone length $\phi_{virtual}(\phi(x_{current}))$ are obtained by calculating. Final real bones $L_{real}$ in the current frame are obtained with attention module. $b$ is the batchsize. $c$ is the number of feature dimension.

Fig. 4. Schematic of virtual bones added to the network. (a) Five bones from pelvis to the joints at the end of the human body. (b) Ten bones between the joints at the end of the human body. (c) 13 bones from pelvis to every nonadjacent joint. (d) 23 bones, the combination of (b) and (c). The joints at the end of the human body are significantly more unstable than others [16], so the virtual bones that input the network are selected. Moreover, considering the human skeleton is based on the root joint (i.e., pelvis), we select the bones related to the four joints at the end of the human body (i.e., head, left-wrist, left-ankle, right-wrist, and right-ankle) and the root joint of the human body.

Finally, four options are determined as shown in Fig. 4. Option one, as shown in Fig. 4(a), contains five virtual bones between the root joint and joints at the end of the human body. Option two, as shown in Fig. 4(b), contains ten virtual bones among joints at the end of the human body. Option three, as shown in Fig. 4(c), contains 13 virtual bones between the root joint and other nonadjacent joints. Option four, as shown in Fig. 4(d), contains 23 virtual bones mentioned in Fig. 4(b) and (c). Experiments are conducted with these options separately to compare their performance. The experiment detail is in Section IV-D.

B. 3-D Joint Prediction

Usually the $k$th joint’s position, $P_k$ is derived as follows:

$$P_k = \sum_{m \in R_k} D_{o,m} \cdot L_m$$  \hspace{1cm} (4)

where $D_{o,m} \in D_o$ and $L_m \in L = L_{real} \cup L_{virtual}$ are the direction and length of bone $m$. $R_k$ is the collection of all bones along the path of the normal human skeleton from the root joint (i.e., pelvis) to the $k$th joint. However, (4) only considers the path along with the real bones. If the virtual bones are added, there will be more than one path to a joint. Hence, the equation changes to

$$P_k = \sum_{R_k, i \in \Lambda_k} w_{k,i} \sum_{m \in R_k, i} D_{o,m} \cdot L_m$$  \hspace{1cm} (5)

where $\Lambda_k$ represents the set of all the paths from the root joint to the $k$th joint. $w_{k,i}$ is the weight of the $i$th path. $R_{k,i}$ is the collection of all the bones on the $i$th path to the $k$th joint.

Thus, in the proposed approach, there are different ways to obtain the position of a joint. A fully connected network is used to calculate the position of joints with the direction and length of both real and virtual bones as input. The network will automatically adjust the weight distribution of every bone related to every predicted joint to obtain the 3-D position. The way obtaining the joints is determined by the bones input, so the selected virtual bones are key of the proposed approach. At the beginning of the experiment, we try the all real bones and virtual bones that between every joints (e.g., a 17-joint-skeleton has 136 bones). However, because the number of bones to be predicted is too large, the time required for each training epoch is too long. In addition, the experimental effect is not good, so we gave up using all the bones.

C. Projection Consistency Loss

In this section, a new loss function, projection consistency loss, is designed. As shown in Fig. 5, considering the position of the 2-D joints in two adjacent frames captured by the same

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
camera, each joint has a certain displacement. Naturally, the 2-D projection of the estimated 3-D joint position will move in the same way as the 2-D input. The calculation process is as follows.

First, according to the pinhole camera model, the estimated 3-D position of each joint is projected back to the 2-D plane

$$\hat{Z}_c \begin{pmatrix} \hat{u} \\ \hat{v} \end{pmatrix} = \begin{pmatrix} \frac{f}{z_c} & 0 & u_0 \\ 0 & \frac{f}{z_c} & v_0 \end{pmatrix} \begin{pmatrix} \hat{X}_j \\ \hat{Y}_j \\ \hat{Z}_c \end{pmatrix}$$

(6)

where $f$ represents the focal length of the camera. $(u_0, v_0)$ represents the position of the optical center of the camera. $(\hat{X}_j, \hat{Y}_j, \hat{Z}_c)$ is the estimated joint coordinates in the camera coordinate system. $(\hat{u}, \hat{v})$ represents the coordinates of the estimated 2-D projection of the joint. $d_x$ and $d_y$ represent zoom factors.

Second, the 2-D displacement of the estimated joint projection is calculated as follows:

$$\Delta_{l,t} = \begin{pmatrix} \hat{u}_{l,t} \\ \hat{v}_{l,t} \end{pmatrix} - \begin{pmatrix} \hat{u}_{l,t-1} \\ \hat{v}_{l,t-1} \end{pmatrix}$$

(7)

where $(\hat{u}_{l,t}, \hat{v}_{l,t})$ represents the $l$th joint’s estimated 2-D projection at time $t$. $\Delta_{l,t}$ means the estimated 2-D displacement of the $l$th joint’s projection from time $t-1$ to time $t$.

Similarly, the ground-truth 2-D displacement is calculated as follows:

$$\Delta_{l,t} = \begin{pmatrix} u_{l,t} \\ v_{l,t} \end{pmatrix} - \begin{pmatrix} u_{l,t-1} \\ v_{l,t-1} \end{pmatrix}$$

(8)

where $(u_{l,t}, v_{l,t})$ represents the $l$th joint’s ground-truth 2-D projection at time $t$. $\Delta_{l,t}$ means the 2-D displacement of the $l$th joint’s ground-truth projection from time $t-1$ to time $t$.

Finally, the projection consistency loss function is given as follows:

$$\text{Loss}_{\text{proj}} = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \left\| \Delta_{j,t} - \Delta_{l,t} \right\|_2$$

(9)

The $\text{Loss}_{\text{proj}}$ can be used to train both the length prediction network and the direction prediction network using estimated joint position at different stages. With the estimated joint position in the bone length prediction stage, $\text{Loss}_{\text{proj-len}}$ is computed to train the length prediction network. $\text{Loss}_{\text{proj-dir}}$ is calculated by the final estimated joint position to train the direction prediction network.

Fig. 6(a) and (b) has the same single-frame projection loss. However, the projection consistency loss of these two situations is very different. Therefore, on the one hand, adding the projection consistency loss can constrain the network and make the estimated joint position more smooth between frames. On the other hand, by constraining the displacement of the joints, the positions of the joints in adjacent frames can be coupled. The joint position of each frame can be constrained in the motion direction of joints from the previous frame to improve the accuracy of the single-frame joint position.

D. Other Loss Functions

The loss function of the bone length prediction network is follows:

$$\text{Loss}_{\text{length}} = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \left\| P_j - \hat{P}_{j,L} \right\|_2$$

(10)

where $\mathcal{J}$ represents the set of all joints. $\hat{P}_{j,L}$ represents the 3-D position of the $j$th joint estimated by bone length estimated network. $P_j$ represents the ground truth of 3-D position of the joint.

The loss function of the bone length attention network is:

$$\text{Loss}_{\text{att}} = \| L - \hat{L} \|_2$$

where $\hat{L}$ represents the length of bones.
estimated by the attention network. \( L \) represents the ground truth of bone lengths.

The loss function of the bone direction prediction network is:
\[
\text{Loss}_{\text{direction}} = \| D - \hat{D} \|_2,
\]
where \( \hat{D} \) represents the direction of bones estimated by the direction prediction network. \( D \) represents the ground truth of bone directions.

The joint shift loss \([14]\) is calculated as follows:
\[
\text{Loss}_{\text{js}} = \sum_{i \in J} \sum_{j \in J, j \neq i} \| (P_i - P_j) - (\hat{P}_i - \hat{P}_j) \|_2 N(i, j) \tag{11}
\]
where \( \hat{P}_i \) represents the position of the \( i \)th joint estimated by the final network.

The loss function of the fully connected layers for the final joint prediction is as follows:
\[
\text{Loss}_{\text{fc}} = \frac{1}{| \mathcal{J} |} \sum_{j \in \mathcal{J}} \| P_j - \hat{P}_j \|_2 \tag{13}
\]
where \( \hat{P}_j \) represents the 3-D position of the \( j \)th joint estimated by the final fully connected layers.

In general, the complete loss function is as follows:
\[
\text{Loss}_{\text{total}} = w_L \text{Loss}_{\text{length}} + w_{\text{att}} \text{Loss}_{\text{att}} + w_d \text{Loss}_{\text{direction}}
+ w_j \text{Loss}_{\text{js}} + w_{pd} \text{Loss}_{\text{proj-dir}} + w_{pl} \text{Loss}_{\text{proj-len}}
+ w_{\text{fc}} \text{Loss}_{\text{fc}} \tag{14}
\]
where \( w_L, w_{\text{att}}, w_d, w_j, w_{pd}, w_{pl}, \) and \( w_{\text{fc}} \) are all hyperparameters.

IV. EXPERIMENTS

A. Data Set and Evaluation

The proposed method is evaluated on the Human3.6M data set \([17]\). Human3.6M provides annotated 2-D and 3-D joint positions of 3.6 million video frames, which contain four camera views for 15 different activities of 11 subjects. Following previous works \([9], [13], [38]\), the training data set is built on five subjects (S1, S5, S6, S7, and S8). The test data set is built on two subjects (S9, S11) with a 17-joint skeleton.

Four protocols are used to evaluate the models. Protocol 1 [mean per-joint position error (MPJPE)] measures the mean Euclidean distance between the predicted and ground-truth joint positions. Protocol 2 (P-MPJPE) is the error between the aligned predicted 3-D joints position and the ground truth. Protocol 3 (N-MPJPE) is the error between the estimated joint position and the ground truth at the same scale. Velocity errors (MPJVE), the errors of the derivative of the corresponding predicted 3-D pose over time, are used to measure the smoothness of the predictions.

B. Implementation Details

The proposed method is tested on the Human3.6M data set, using the 2-D coordinates of the cascaded pyramid network (CPN) \([9]\) or the 2-D coordinates of ground truth as the model input. The visibility score used in the proposed method comes from AlphaPose \([41]\). The results of the baseline method \([14]\) are experimented based on their open-source codes.

The optimizer of the network is Adam \([42]\). The batchsize \( b \) is 2048 for the 9-frame model (9-frame receptive field) and 1024 for the 243-frame model (243-frame receptive field). The number of training epochs is 60. Learning rate is set to 0.001 and the learning rate decays at the rate of 0.95 per epoch. We set \( w_{\text{proj-dir}} = 1 \) and \( w_{\text{proj-len}} = 1 \) for the total loss function. The other hyperparameters are the same as \([14]\), such as \( w_d = 0.02, w_j = 1, w_{\text{att}} = 0.05, w_j = 0.1, \) and \( w_{\text{fc}} = 1 \).

In addition, because the bone direction prediction network can only give the bone direction estimation for the middle frame for each video, the middle frames of different videos under the same camera are used to replace the adjacent frames to calculate \( \text{Loss}_{\text{proj-dir}} \). Three NVIDIA 1080Ti GPUs are used to train the 243-frame model and one for the 9-frame model.

C. Experimental Results

Tables I and II show the quantitative comparison of the accuracy between the proposed method and other existing methods on the Human3.6M data set. For using CPN \([9]\) 2-D inputs in Table I, our method achieves performance similar to state-of-the-art methods when using a large model with the 243-frame receptive fields. However, for using 2-D truth values as input, our method outperforms the state-of-the-art method \([14]\) on all estimation protocols. Therefore, our slightly worse performance when using CPN \([9]\) 2-D inputs is due to errors of the input information. When using accurate information as input, that is, 2-D joint ground truth, our method shows high performance. Under the experimental condition of 9-frame receptive fields, the proposed method gets better results than \([14]\) both with CPN \([9]\) 2-D inputs and the 2-D ground-truth inputs. The good performance of the proposed method using a smaller model means less computational resources and time consumption. Fig. 7 is the visualized results of the proposed method in the two actions: 1) phoning and 2) walking dog.

Table III shows the parameter sensitivity of the proposed method. Only the parameters we introduced are tested. The proposed method is sensitive to the choice of hyperparameters, so their values are set based on the test results in the table.

In addition, we also tested the time statistics of the proposed method. In our method, we have a lightweight model (9 frame) and a highly accurate model (243 frame). In recent years, most videos for human pose estimation are recorded at a frequency of 25 Hz, so in the real-time test, we believe that the prediction time of 25 frames is less than 1 s to meet the real-time performance. For the 9-frame model, the time to predict 25 frames is 0.54 s, less than 1 s. For the 243-frame model, the time is 1.19 s, more than 1 s. The lightweight model meets the real-time requirement.

D. Ablation Study

The ablation experiments are performed on Human3.6M under Protocols 1, 2, and 3, and velocity. The 9-frame models and 243-frame models are used, respectively, for the comparisons between the baseline \([14]\) and the proposed method.
### TABLE I

**Comparisons of the Proposed Method With Other Existing Methods in All Actions of the Human 3.6M Data Set Under the Metrics Protocol 1, Protocol 2, and MPJVE. The Results Are Based on 2-D Joint Input From CPN [9].**

**a)** Under the Metric Protocol 1. **b)** Under the Metric Protocol 2. **c)** Under the Metric MPJVE.

| Methods     | Dist | Ect | Groat | Photo | Pmc | Sih | Sih | Tsmk | Wait | WalkD | WalkT | Avg |
|-------------|------|-----|-------|-------|-----|-----|-----|------|------|------|-------|-----|
| Martinez et al. [34] | 51.8 | 56.2 | 58.1 | 60.9 | 69.5 | 78.4 | 55.2 | 58.1 | 74.0 | 94.6 | 62.3 | 59.1 |
| Sun et al. [35] | 53.8 | 54.8 | 54.2 | 54.3 | 61.8 | 67.2 | 53.1 | 53.6 | 71.7 | 86.7 | 61.5 | 53.4 |
| Pavlakos et al. [66] | 48.5 | 54.4 | 54.4 | 59.4 | 65.4 | 54.8 | 53.6 | 57.5 | 73.9 | 41.5 | 47.9 | 51.2 |
| Yang et al. [37] | 51.5 | 58.9 | 50.4 | 57.0 | 62.1 | 65.1 | 49.8 | 52.7 | 69.2 | 85.2 | 57.4 | 58.4 |
| Liu et al. [12] | 49.2 | 51.6 | 47.6 | 50.5 | 51.8 | 60.3 | 48.5 | 51.7 | 61.5 | 70.9 | 53.7 | 48.9 |
| Hossain & Little [38] | 48.0 | 50.7 | 52.2 | 55.2 | 63.1 | 72.6 | 53.0 | 51.7 | 66.1 | 80.3 | 59.0 | 57.3 |
| Lee et al. [13] | 40.2 | 49.2 | 47.8 | 52.6 | 50.1 | 75.0 | 50.2 | 43.0 | 55.8 | 73.9 | 54.1 | 55.6 |
| Chen et al. [10] | 41.1 | 44.2 | 44.9 | 45.9 | 46.5 | 39.3 | 41.6 | 34.8 | 73.3 | 46.2 | 48.7 | 42.1 |
| Pavlakos et al. [9] (243 frames) | 45.9 | 48.5 | 46.3 | 47.8 | 51.9 | 57.5 | 46.2 | 45.6 | 59.9 | 68.5 | 50.6 | 46.4 |
| Pavlakos et al. [9] (243 frames) | 45.2 | 46.7 | 43.3 | 45.6 | 48.1 | 55.1 | 44.6 | 44.3 | 55.3 | 47.8 | 41.0 | 49.0 |
| Liu et al. [11] (243 frames) | 45.2 | 44.8 | 42.6 | 46.5 | 48.3 | 59.7 | 41.4 | 44.6 | 56.4 | 47.4 | 41.5 | 53.7 |
| Cai et al. [9] (243 frames) | 44.6 | 46.5 | 45.8 | 48.8 | 50.0 | 47.2 | 42.9 | 57.9 | 61.9 | 47.6 | 51.3 | 57.1 |
| Ye et al. [40] (NPS19) | 44.8 | 46.1 | 43.3 | 46.4 | 49.0 | 55.2 | 44.6 | 44.0 | 58.3 | 62.7 | 47.1 | 43.9 |
| Xu et al. [16] (243 frames) | 37.4 | 43.5 | 42.7 | 42.7 | 46.6 | 59.7 | 41.8 | 48.1 | 52.7 | 60.2 | 45.8 | 41.1 |
| Chen et al. [14] (9 frames) | 44.4 | 47.0 | 43.2 | 46.6 | 46.0 | 57.1 | 40.6 | 44.0 | 55.9 | 61.1 | 48.5 | 40.5 |
| Chen et al. [14] (243 frames) | 41.4 | 43.0 | 40.1 | 42.9 | 46.6 | 51.9 | 41.7 | 42.3 | 53.9 | 60.2 | 43.4 | 41.7 |
| ours (9 frames) | 44.4 | 46.1 | 44.3 | 46.2 | 46.9 | 57.1 | 45.2 | 43.9 | 55.6 | 61.2 | 48.5 | 45.6 |
| ours (243 frames) | 42.4 | 43.5 | 41.0 | 43.5 | 46.7 | 54.6 | 42.5 | 42.1 | 54.0 | 65.7 | 42.1 | 46.5 |

### TABLE II

**Results of 9-Frame and 243-Frame Models on Human 3.6M With the Ground-Truth 2-D Input**

1) Influence of Numbers of Virtual Bones: According to the comparison of the results in Table IV(a), when five virtual bones are added, most protocols are slightly improved. Too few virtual bones added to the joint prediction can only bring a slight improvement. When 10, 13, and 23 virtual bones are added, all evaluation protocols have great improvement compared to the baseline. In a smaller receptive field, the proposed method has produced better results. Our method can have a greater improvement when there is less information input, which shows that our method is more effective when there is more room for improvement.

2) Influence of Projection Consistency Loss: It can be seen from Table IV(b) that although increasing the projection consistency loss alone has little effect on the 243-frame model, it can effectively improve the performance of baseline under the 9-frame receptive field. Another point worth paying attention to is that increasing the projection consistency loss alone has a greater improvement on Protocols 1 and 3, but the impact on Protocol 2 is more limited. Protocol 2 is obtained by calculating the minimum error of the skeleton after the rigid-body transformation, so the accuracy of the joint position after rotation is less improved.

### TABLE III

**Parameter Sensitivity Test of 9-Frame Model on the Human 3.6M Data Set, Using Ground-Truth 2-D Joint Positions as Inputs.**

| Methods | Protocol 1 | Protocol 2 | Protocol 3 | MPJVE |
|---------|------------|------------|------------|-------|
| Martinez et al. [34] | 45.7 | - | - | - |
| Hossain & Little [38] | 41.6 | - | - | - |
| Lee et al. [13] | 38.4 | - | - | - |
| Pavlakos et al. [9] (243 frames) | 37.2 | 27.2 | - | - |
| Chen et al. [14] (9 frames) | 38.0 | 28.2 | 37.3 | 1.96 |
| Others (9 frames) | 35.4 | 27.2 | 34.7 | 1.92 |
| Chen et al. [14] (243 frames) | 38.0 | 29.7 | 34.9 | 1.94 |
| Others (243 frames) | 32.5 | 25.2 | 31.9 | 1.70 |

*Except for the conditions to be compared, other experimental settings are the same as Section IV-B.*

1) Influence of Numbers of Virtual Bones: According to the comparison of the results in Table IV(a), when five virtual bones are added, most protocols are slightly improved. Too few virtual bones added to the joint prediction can only bring a slight improvement. When 10, 13, and 23 virtual bones are added, all evaluation protocols have great improvement compared to the baseline. In a smaller receptive field, the proposed method has produced better results. Our method can have a greater improvement when there is less information input, which shows that our method is more effective when there is more room for improvement.

2) Influence of Projection Consistency Loss: It can be seen from Table IV(b) that although increasing the projection consistency loss alone has little effect on the 243-frame model, it can effectively improve the performance of baseline under the 9-frame receptive field. Another point worth paying attention to is that increasing the projection consistency loss alone has a greater improvement on Protocols 1 and 3, but the impact on Protocol 2 is more limited. Protocol 2 is obtained by calculating the minimum error of the skeleton after the rigid-body transformation, so the accuracy of the joint position after rotation is less improved.
Fig. 7. Visualized comparison of the results of the baseline and the proposed method, both of which use a 9-frame receptive field and are trained on the CPN [9] 2-D inputs. As shown in the red mark in the picture, our prediction has higher accuracy in the end joints due to the application of virtual bones. In addition, our method also performs better under complex human pose and large motions due to the usage of motion constraints based on projection consistency loss.

TABLE IV

| Comparison of Different Models Under Protocols on Human3.6M. “Baseline” Represents the Baseline 9-Frame and 243-Frame Models We Experiment Based on [14]. Other Rows Represent the Loss or Virtual Bones We Proposed. They Are Used, Respectively, to Test the Impact on the Baseline Model. “5VB, 10VB, 13VB, and 23VB” Refer to the Different Numbers of Virtual Bones Selected to Be Added to the Bone Prediction Network. “PCL” Refers to the Projection Consistency Loss. |  |
|---|---|---|---|---|---|---|---|---|---|
| | 9-frame | | | | 243-frame | | | | |
| | Protocol 1 | Protocol 2 | Protocol 3 | Velocity | Protocol 1 | Protocol 2 | Protocol 3 | Velocity |
| Baseline | 38.0 | 28.2 | 37.3 | 1.96 | 34.0 | 25.9 | 33.3 | 1.71 |
| Baseline+5VB | 37.1 | 27.9 | 36.0 | 1.94 | 34.1 | 25.6 | 33.4 | 1.70 |
| Baseline+10VB | 37.1 | 27.9 | 36.0 | 1.96 | 33.4 | 25.4 | 32.8 | 1.70 |
| Baseline+13VB | 36.7 | 27.4 | 35.8 | 1.93 | 33.5 | 25.4 | 32.6 | 1.69 |
| Baseline+23VB | 36.7 | 27.4 | 35.3 | 1.94 | 33.5 | 25.3 | 32.7 | 1.69 |
| (b) Baseline+PCL | 36.8 | 27.9 | 36.0 | 1.94 | 34.1 | 26.2 | 33.6 | 1.70 |

V. Conclusion

In this article, a novel 3-D human pose prediction network and a novel projection consistency loss are proposed. Virtual bones between nonadjacent joints are proposed to optimize the estimation of bone length. Random frames are used to predict the real bone length combined with an attention...
mechanism, and the current frame is used to predict virtual bone length directly. The bone direction prediction network is implemented by a temporal convolutional network to predict direction. Moreover, a 2-D projection consistency loss is presented to constrain the motion displacement of joints between adjacent frames. Experiments indicated that the improved framework performs well in the 9-frame receptive field. The study of graphs composed of real bones and virtual bones based on graph networks will be our future direction.

REFERENCES

[1] S. A. W. Talha, M. Hammouche, E. Ghorbel, A. Fleury, and S. Amblbousis, “Features and classification schemes for view-invariant and real-time human action recognition,” *IEEE Trans. Cogn. Devel. Syst.*, vol. 10, no. 4, pp. 894–902, Dec. 2018.

[2] N. T. V. Tuyen, A. Elihou, and N. Y. Chong, “Learning bodily expression of emotion for social robots through human interaction,” *IEEE Trans. Cogn. Devel. Syst.*, vol. 13, no. 1, pp. 16–30, Mar. 2021.

[3] G. Wang, Y. Yang, H. Zhang, Z. Liu, and H. Wang, “Spherical interpolated convolutional network with distance-feature density for 3-D semantic segmentation of point clouds,” *IEEE Trans. Cybern.*, early access, Nov. 8, 2021, doi: 10.1109/TCYB.2021.3124954.

[4] Z. Liu, Z. Qiao, C. Suo, Y. Liu, and K. Jin, “Map-less long-term localization in complex industrial environments,” *Assam. Autom.*, vol. 41, no. 6, pp. 714–724, 2021.

[5] A. Zaraki et al., “Design and evaluation of a unique social perception system for human–robot interaction,” *IEEE Trans. Cogn. Devel. Syst.*, vol. 9, no. 4, pp. 341–355, Dec. 2017.

[6] K. Li, H. Wang, X. Liang, and Y. Miao, “Visual Servoing of flexible-link manipulators by considering vibration suppression without deformation,” *IEEE Trans. Cybern.*, early access, May 26, 2021, doi: 10.1109/TCYB.2021.3072779.

[7] Z. Liu, H. Wei, H. Wang, H. Li, and H. Wang, “Integrated task allocation and path coordination for large-scale robot networks with uncertainties,” *IEEE Trans. Autom. Sci. Eng.*, early access, Sep. 23, 2021, doi: 10.1109/TASE.2021.3111888.

[8] M. Xu, A. Hu, and H. Wang, “Image-based visual impedance force control for contact manipulation,” *IEEE Trans. Autom. Sci. Eng.*, early access, Apr. 6, 2022, doi: 10.1109/TASE.2022.3162207.

[9] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, “3D human pose estimation via graph convolutional networks,” in Proc. *IEEE/CVF Int. Conf. Comput. Vis.*, 2020, pp. 5243–5252.

[10] S. Li, L. Ke, K. Pratama, Y.-W. Tai, C.-K. Tang, and K.-T. Cheng, “Cascade deep monocular 3D human pose estimation with evolutionary training data,” in Proc. *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6173–6183.

[11] R. Liu, J. Shen, H. Wang, C. Chen, S.-C. Cheung, and V. Asari, “Attention mechanism exploits temporal contexts: Real-time 3D human pose reconstruction,” in Proc. *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5064–5073.

[12] Y. Cheng, B. Yang, B. Wang, and R. T. Tan, “3D human pose estimation using spatio-temporal networks with explicit occlusion training,” in Proc. *AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 10631–10638.

[13] H. Wu and B. Xiao, “3D human pose estimation via explicit compositional depth maps,” in Proc. *AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 12378–12385.

[14] M. Jiang, Z. Yu, Y. Zhang, Q. Wang, C. Li, and Y. Lei, “Reweighted sparse representation with residual compensation for 3D human pose estimation from a single RGB image,” *Neurocomputing*, vol. 358, pp. 332–343, Sep. 2019.

[15] F. Moreno-Noguer, “3D human pose estimation from a single image via distance matrix regression,” in Proc. *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2823–2832.

[16] X. Guo and Y. Dai, “Occluded joints recovery in 3D human pose estimation based on distance matrix,” in Proc. *24th Int. Conf. Pattern Recognit. (ICPR)*, 2018, pp. 1325–1330.

[17] X. Guo, W. Hu, J. Tang, J. Liu, and Z. Guo, “Optimized skeleton-based action recognition via sparsified graph regression,” in Proc. *27th ACM Int. Conf. Multimedia*, 2019, pp. 601–610.

[18] J. Martinez, R. Hossain, J. Romero, and J. J. Little, “A simple yet effective baseline for 3D human pose estimation,” in Proc. *IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2640–2649.

[19] X. Sun, J. Shang, S. Liang, and Y. Wei, “Compositional human pose regression,” in Proc. *IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2602–2611.

[20] G. Pavlakos, X. Zhou, and K. Daniilidis, “Ordinal depth supervision for 3D human pose estimation,” in Proc. *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7307–7316.

[21] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, “3D human pose estimation in the wild by adversarial learning,” in Proc. *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5255–5264.

[22] M. R. I. Hossain and J. J. Little, “Exploiting temporal information for 3D human pose estimation,” in Proc. *Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 68–84.

[23] Y. Cai et al., “Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks,” in Proc. *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 2272–2281.

[24] R. Yeh, Y.-T. Hu, and A. Schwing, “Chirality nets for human pose regression,” in *Advances in Neural Information Processing Systems*, vol. 32, Red Hook, NY, USA: Curran Assoc., 2019.

[25] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, “RMPE: Regional multi-person pose estimation,” in Proc. *IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2334–2343.

[26] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.
Guangming Wang (Graduate Student Member, IEEE) received the B.S. degree from the Department of Automation, Central South University, Changsha, China, in 2018. He is currently pursuing the Ph.D. degree in control science and engineering with Shanghai Jiao Tong University, Shanghai, China.

His current research interests include computer vision and SLAM, in particular, 3-D human pose estimation.

Honghao Zeng is currently pursuing the B.S. degree with the Department of Automation, Shanghai Jiao Tong University, Shanghai, China.

His current research interests include computer vision and SLAM, in particular, 3-D human pose estimation.

Ziliang Wang is currently pursuing the B.S. degree with the Department of Automation, Shanghai Jiao Tong University, Shanghai, China.

His current research interests include computer vision and SLAM, in particular, 3-D human pose estimation.

Zhe Liu received the B.S. degree in automation from Tianjin University, Tianjin, China, in 2010, and the Ph.D. degree in control technology and control engineering from Shanghai Jiao Tong University, Shanghai, China, in 2016.

From 2017 to 2020, he was a Postdoctoral Fellow with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong. He is currently a Research Associate with the Department of Computer Science and Technology, University of Cambridge, Cambridge, U.K. His research interests include autonomous mobile robot, multirobot cooperation, and autonomous driving system.

Hesheng Wang (Senior Member, IEEE) received the B.Eng. degree in electrical engineering from Harbin Institute of Technology, Harbin, China, in 2002, and the M.Phil. and Ph.D. degrees in automation and computer-aided engineering from The Chinese University of Hong Kong, Hong Kong, in 2004 and 2007, respectively.

He is currently a Professor with the Department of Automation, Shanghai Jiao Tong University, Shanghai, China. His current research interests include visual servoing, service robot, computer vision, and autonomous driving.

Dr. Wang is an Associate Editor of IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, IEEE ROBOTICS AND AUTOMATION LETTERS, Assembly Automation, and the International Journal of Humanoid Robotics; a Technical Editor of the IEEE/ASME TRANSACTIONS ON MECHATRONICS; and an Editor of Conference Editorial Board of IEEE Robotics and Automation Society. He served as an Associate Editor of the IEEE TRANSACTIONS ON ROBOTICS from 2015 to 2019. He was the General Chair of IEEE ROBIO 2022 and IEEE RCAR 2016, and the Program Chair of the IEEE ROBIO 2014 and IEEE/ASME AIM 2019. He will be the General Chair of IEEE/RSJ IROS 2025.