Unsupervised 3D Human Pose Representation with Viewpoint and Pose Disentanglement

Qiang Nie[1,2][0000−0002−2778−4058], Ziwei Liu[1][0000−0002−4220−5958], and Yunhui Liu[1,2][0000−0002−3625−6679]

[1] The Chinese University of Hong Kong, Shatin N.T., Hong Kong
[2] T Stone Robotics Institute of CUHK
{qnie,yhliu}@mae.cuhk.edu.hk, zwliu@ie.cuhk.edu.hk

Abstract. Learning a good 3D human pose representation is important for human pose related tasks, *e.g.* human 3D pose estimation and action recognition. Within all these problems, *preserving the intrinsic pose information* and *adapting to view variations* are two critical issues. In this work, we propose a novel Siamese denoising autoencoder to learn a 3D pose representation by disentangling the pose-dependent and view-dependent feature from the human skeleton data, in a fully unsupervised manner. These two disentangled features are utilized together as the representation of the 3D pose. To consider both the kinematic and geometric dependencies, a sequential bidirectional recursive network (SeBiReNet) is further proposed to model the human skeleton data. Extensive experiments demonstrate that the learned representation 1) preserves the intrinsic information of human pose, 2) shows good transferability across datasets and tasks. Notably, our approach achieves state-of-the-art performance on two inherently different tasks: pose denoising and unsupervised action recognition. Code and models are available at: `https://github.com/NIEQiang001/unsupervised-human-pose.git`.

Keywords: Representation Learning, 3D Human Pose, Pose Denoising, Unsupervised Action Recognition.

1 Introduction

Human action recognition and human behavior analysis have extensive applications on human-robot interaction (HRI) systems, such as health caring, entertainment, education, security and many other intelligent surveillance scenarios, which also makes the 3D human pose estimation a hot research topic for many decades. Learning a good human 3D pose representation has great significance both to the research of human action recognition and the human pose estimation.

While understanding the human pose is a challenging task, which requires the computer to learn the dependencies between joints of the human skeleton robustly in different viewpoints. These dependencies include kinematic relationships between joints and geometric features of the human body. The kinematic relationship describes the motion transmission process between joints and the role of each joint in an action. The geometric feature refers to those specific appearance characteristics of the human body, such as fixed bone lengths and the symmetry between left and right limbs. Many existing works have utilized the geometric features of the human body [17,13,22,36,18], but few works are capable to model the kinematic relationships between human body joints. Kinematics is a physical process and hard to be modeled by regular CNN, RNN or MLP neural networks. Hence, we proposed a sequential bidirectional recursive network (SeBiReNet) to model the dependencies of the human skeleton.

Besides the dependencies between joints, the human 3D pose presents infinite modalities when recorded or observed from different viewpoints, which makes the processing of the human 3D pose quite intractable for the intelligent system. Increasing the size of training dataset with different views may be effective. However, it's impossible to record the data from all possible viewpoints. To tackle the view variation, some previous works applied preprocessing treatment to eliminate the view variation [12,3]. These methods are always dataset dependent because of the specifically designed preprocessing method. Many other methods [32,16,32,6,23,29] extracted hand-crafted view-invariant features as pose descriptors based on the prior knowledge of human beings. Although these hand-crafted features are view-invariant, there is information loss in extracting these features as only a few explanatory factors are considered. There are some methods [3,9,35] trying to learn discriminative pose representations using the deep learning method. However, the transferability of the representations learned by existing approaches in different datasets and different tasks is limited.

Human pose result from the rich interaction of many factors, such as the subject, the action, and the viewpoint. Learning view-invariant features means to extract features that are insensitive to the direction of view variation, which also means some features that are sensitive to the variations but informative are discarded. As Bengio et al. [2] mentioned, a better way to overcome these challenges is to leverage the data itself, ..., to disentangle as many factors as possible and discarding as little information about the data as in practice.

Motivated by aforementioned issues, we propose an unsupervised method for learning a latent representation of the human 3D pose by disentangling the pose-dependent and view-dependent features from human skeleton data. We introduce a novel SeBiReNet to model the human skeleton data. A Siamese denoising autoencoder based on the SeBiReNet is designed to learn the latent human pose representation. Ability of denoising corrupted skeletons from an unseen dataset proves the learned representation preserves the intrinsic information of human pose, including both the kinematic and geometric dependencies. Disentangling the pose-dependent (view-invariant) and view-dependent (view-variant) feature from skeleton data other than extracting the view-invariant feature enables us to transfer the viewpoint of human pose in the latent space, which is used as a strengthened regularization in our training process.

We summarize our contributions as follows:

- We propose a novel SeBiReNet to model the kinematic dependencies between body joints in the human skeleton data.

- Based on SeBiReNet, a Siamese denosing autoencoder is proposed for learning 3D human pose representation with feature disentanglement. The unsupervisedly learned pose representation 1) preserves the intrinsic information of human pose, 2) shows good transferability across datasets and tasks.
- Extensive experiments demonstrate that state-of-the-art performance can be achieved when applying the learned representation on two inherently different tasks: pose denoising and unsupervised action recognition.

2 Related Works

2.1 Modeling Human 3D Poses

To understand the human 3D pose, the most important is to figure out the dependencies between body joints, which should include both the kinematic and the geometric dependency. Compared to kinematic dependency between joints, geometric characteristic is much easier to model. Ramakrishna et al. [17] used normalized limb lengths as anthropometric regularity to reconstruct 3D human pose from the 2D image landmarks. Sun et al. [22] proposed to use the summation of bone lengths as a supervision loss. The summation of bone length considers all bones between every paired two joints. In essence, the summation of bone lengths is a pairwise geodesic distance. Their work proved that the accuracy of human pose regression can be improved based on the summations of bone length. As ratios between bone lengths remain relatively fixed in a human skeleton, Zhou et al. [36] utilized the length ratios of bones as a weak supervision for reconstructing 3D pose from wild images without 3D annotations. Though human skeleton is similar to the tree structure, few works have applied the recursive network for the human pose modeling. Wei et al. [28] introduced a vanilla tree network for skeleton-based action recognition. However, only the output from the single tree root node is utilized, which is inherently different from the structure of our SeBiReNet proposed to model the human 3D pose.

2.2 Learning Pose Representations

Demisse et al. [3] proposed a denoising autoencoder for unsupervised skeleton-based action recognition by using MLP layers. But to evaluate the extracted features in cross-view action recognition, a preprocessing treatment is applied to estimate the view variation. Li et al. [9] proposed a method to learn pose representation from sequential RGB data by adding a view discriminator to decide which view the learned feature comes from. Using view classifier indicates that their views are depend on the training data and view labels were given. While in our method, the poses are randomly rotated and no view label is given. Zheng et al. [35] presented an adversarial training strategy to learn representations of skeleton sequences for action recognition. Compared to these methods, the proposed method is able to learn a view-invariant pose-dependent feature from single pose without any additional label or auxiliary network. Requiring no temporal information makes our representation can be applied to both time-related or time-independent tasks, as verified in our experiments. It's interesting to find that Aberman et al. [1] applied a similar feature decomposition and re-composition process in their work of retargeting video-captured motion between different human performers. Our method differs with theirs in two aspects: 1) we embed a denoising process in the learning, which helps the network capture the intrinsic feature of skeleton pose; and 2) our disentangled features have more interaction by sharing some weights in the decomposition and multiplying with each other in the re-composition process.

3 Our Approach

3.1 Problem Formulation

Given a human 3D pose x, a latent representation h can be learned by assigning a function f with parameters θ, i.e., $h = f(x;\theta)$. In order to make sure the learned representation contains useful information of original data, h is required to be able to recover the original pose through another function g with parameters ζ. The reconstructed pose can be formulated as $\hat{x} = g(h;\zeta)$, which is the basic idea of autoencoder in representation learning. Vincent et al. [24] has proven that using the denoising autoencoder to reconstruct the clean data from its corrupted version is helpful in avoiding trivial solutions and improving the performance of learned latent representations. Therefore, basically, learning the human 3D pose representation can be formulated as the following equation.

$$\arg\min_{\theta,\zeta} \mathbb{E}_{q(\tilde{x}|x)} L(x, g(h;\zeta)) \tag{1}$$

where $h = f_\theta(\tilde{x})$ is the learned latent representation of the human 3D pose and $\tilde{x}$ is the corrupted pose corresponding to the clean pose x. However, the representation learned in eq. 1 contains both the pose-dependent and the view-dependent information. As analyzed in Sec. 1, we hope to learn a view-invariant representation as well as avoid discarding the view-dependent feature of human pose for the sake of information preservation. Thus different from traditional methods, we attempt to disentangle the view-invariant feature h_{vi} from view-dependent feature h_v and using the combination of $[h_{vi}, h_v]$ as a latent representation of the human 3D pose. Under this consideration, the representation learning is reformulated as eq. 2.

$$\arg\min_{\theta_{vi},\theta_v,\zeta} \mathbb{E}_{q(\tilde{x}|x)} L(x, g(h_v \otimes h_{vi};\zeta)) \tag{2}$$

where $h_{vi} = f(\tilde{x};\theta_{vi})$ and $h_v = f(\tilde{x};\theta_v)$. $\otimes$ is an operation to couple h_{vi} and h_v together, which can be matrix multiplication or concatenation. From a generative perspective, the learning process in eq. 2 can also be written as

$$\arg\max_{\theta_{vi},\theta_v,\zeta} \mathbb{E}_{q(\tilde{x}|x)} \log\left[p(x|h_{vi},h_v;\zeta)p(h_{vi},h_v|\tilde{x};\theta_{vi},\theta_v)p(\theta_{vi},\theta_v)\right] \tag{3}$$

where $q(\tilde{x}|x)$ denotes the pose corruption process. If we assume the prior distribution $p(\theta_{vi}, \theta_v)$ can be factorized as $p(\theta_{vi}, \theta_v) = p(\theta_{vi})p(\theta_v)$, i.e., they are independent. Then we have

$$\log p(h_{vi}, h_v|\tilde{x}; \theta_{vi}, \theta_v)p(\theta_{vi}, \theta_v) = \log p(h_{vi}|\tilde{x}; \theta_{vi})p(\theta_{vi}) + \log p(h_v|\tilde{x}; \theta_v)p(\theta_v) \tag{4}$$

According to eq. 4, learning of the view-dependent feature and pose-dependent feature don't have much influence on each other. To strengthen the interaction between these two features and disentangle them smoothly, we propose to have $p(h_{vi}|\tilde{x}; \theta_{vi}) = f(\phi(\tilde{x}; \eta); \theta_{vi} \backslash \eta)$ and $p(h_v|\tilde{x}; \theta_v) = f(\phi(\tilde{x}; \eta); \theta_v \backslash \eta)$, where η is the shared parameters in the parameter space. In this manner, h_v and h_{vi} are disentangled and affect each other through the common latent feature $\phi(\tilde{x})$. Although we are trying to disentangle the view-dependent and pose-dependent feature from original pose, this is not necessarily induced so far, as the learned latent representation $[h_{vi}, h_v]$ is still not well constrained. To introduce the concept of viewpoint into the learning process, an additional transformation is added to the corrupted pose $\tilde{x}$ by randomly rotating it in the 3D space. At this circumstance, eq. 3 becomes

$$\arg\max_{\theta_{vi}, \theta_v, \zeta} \mathbb{E}_{q(\tilde{x}|x)q_r(\tilde{x}_r|\tilde{x})} \log\left[p(x|h_{vi}, h_v; \zeta)p(h_{vi}, h_v|\tilde{x}_r; \theta_{vi}, \theta_v)p(\theta_{vi}, \theta_v)\right] \tag{5}$$

where $\tilde{x}_r$ is the randomly rotated corrupted pose corresponding to $\tilde{x}$, $q_r(\tilde{x}_r|\tilde{x})$ denotes the random rotation process. A marginal distribution consistency of $p(h_{vi})$ should be satisfied from corrupted pose $\tilde{x}$ and randomly rotated pose $\tilde{x}_r$. Thus, besides the pose reconstruction loss, we regularize the pose-dependent feature by minimizing the Kullback–Leibler divergence between pose-dependent features of poses under different observation angles as shown in eq. 6.

$$\arg\min_{\theta_{vi}, \theta_v} D_{KL}(p(h_{vi}|\tilde{x}; \theta_{vi})||p(h_{vi}|\tilde{x}_r; \theta_{vi})) \tag{6}$$

Putting all together, our human 3D pose representation learning process is modelled as

$$\arg\min_{\theta_{vi}, \theta_v, \zeta} \mathbb{E}_{q(\tilde{x}|x)}\left[L(x, g(h_v \otimes h_{vi}, \tilde{x}; \zeta)) + q_r(\tilde{x}_r|\tilde{x})L(x, g(h_v \otimes h_{vi}, \tilde{x}_r; \zeta))\right] + D_{KL}(p(h_{vi}|\tilde{x}; \theta_{vi})||p(h_{vi}|\tilde{x}_r; \theta_{vi})) \tag{7}$$

3.2 Sequential Bidirectional Recursive Network

In order to capture the kinematic dependencies of human skeleton structure, a sequential bidirectional recursive neural network (SeBiReNet) is proposed. The bidirectional recursive neural network has two tree structures as shown in Fig. 1, which models the human skeleton structure intuitively. The recursive neural network is widely used for text or language analysis [7,21] due to its ability in summarising the semantic meanings. However, the conventional recursive neural network has only one direction, which means the information can only

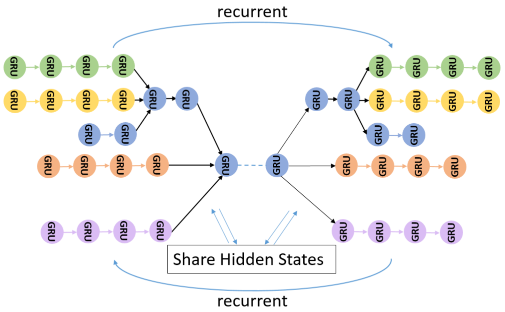

Fig. 1. The proposed sequential bidirectional recursive neural network (SeBiReNet). Each node corresponds to a real joint of the human body and different colors represent different body parts

flow from leaf nodes to the root node. On the contrary, the motion of the human body is transmitted from parent joint to child joints. Usually, to determine the position of a joint, both the position of parent joint and the positions of child joints have to be considered. In this regards, the proposed SeBiReNet models the dependency $p(J_{parent}|J_{child})$ and $p(J_{child}|J_{parent})$ between parent joint and child joint respectively through a recursive subnet (left part in Fig. 1) and a diffuse subnet (right part in Fig. 1). The two subnets have independent kernel weights but share the hidden states $h \in \mathbb{R}^{J \times m}$, where J is the joint number and m is the feature size. The shared hidden states store the intermediate inference results when information flows in the network, and the intermediate results will be continually refined when the network recurrently runs. This network is named SeBiReNet because information flows sequentially and reversely in the two subnets. The proposed architecture not only models the forward and inverse kinematic process but also imitates the repeated thinking process of human.

The node number of SeBiReNet can be adjusted according to the joint number of a human skeleton model. As most skeleton models contain 17 joints, the basic version of our proposed model is designed to have 34 nodes. In SeBiReNet, each node is a GRU cell. Other node types, such as LSTM, can also be used. The forgetting mechanism GRU cell enables the network to tackle noisy input. The inference process of SeBiReNet can be formulated as equation 8.

$$\begin{aligned}
h_i^r &= \varphi(W_{xi}^r x_i^r + W_{hi}^r h_i + b_i^r) \\
O_i^r &= \mathcal{O}(W_o^r h_i^r + b_o^r) \\
h_i^d &= \varphi(W_{xi}^d x_i^d + W_{hi}^d h_i + b_i^d) \\
O_i^d &= \mathcal{O}(W_o^d h_i^d + b_o^d)
\end{aligned} \tag{8}$$

where $x_i^r = (p_i, h_{children})$ and $x_i^d = (p_i, h_{parent})$ are the input of the node i, which contains the 3D position p_i of corresponding joint i and the hidden states output from all its child nodes $h_{children}$ or parent node h_{parent}. $h_i \in \mathbb{R}^m$ denotes the shared hidden state of the node i. The superscript r represents the recursive subnet and $W^r_{xi}, W^r_{hi}, b^r_i, W^r_o, b^r_o$ are kernel weights and biases of it. The superscript d denotes the diffuse subnet and $W^d_{xi}, W^d_{hi}, b^d_i, W^d_o, b^d_o$ are kernel weights and biases belong to it. φ denotes the nonlinear function of the GRU cell. $\mathcal{O}$ is the activation function and $tanh$ is used in our experiments. After each inference in the recursive subnet or diffuse subnet, the shared hidden states and network output will be updated by the hidden states and output of corresponding subnet. Outputs of all nodes are concatenated together as the final output of the SeBiReNet.

Complexity of the SeBiReNet Assuming the hidden units of each GRU node is n_h and the dimension of input feature is n_x. The number of parameters in a node with N child nodes (in recursive subnetwork) or parent node (in diffuse subnetwork) is $l_N = [3n_x + (3N+4)n_h + 1]\, n_h$. In a SeBiReNet with 17 joints, there are 6 leaf nodes (l_0), 26 nodes with one child node or parent node (l_1), 2 nodes have 3 child nodes (l_3).

3.3 Learning Framework Based on SeBiReNet

According to the analysis in Sec. 3.1, we designed a denoising autoencoder (DAE) to learn the representation of the human 3D pose based on SeBiReNet. Different from general practice [25] that adds Gaussian noise to the clean input and achieves a gently polluted version, we directly destroy the skeleton to an unreasonable version where some randomly selected joints are moved to illegal positions. The network is expected to distinguish valid human pose from invalid human pose and recover the correct position of those invalid joints.

Though the kinematic dependency has intrinsically modeled by the SeBiReNet, the geometric characteristics haven't been well considered. To this end, we added a bone length loss L_B and a symmetry loss L_S to the pose reconstruction loss, as shown in eq. 9.

$$L(p, g(h_v \otimes h_{vi}, \tilde{p}; \zeta)) = \sum_s (L_P + L_B + L_S) \tag{9}$$

where the first part $L_P = \sum_{i=1}^{J} \|p^s_i - \hat{p}^s_i)\|_2$ is the reconstruction error of joint position, p^s_i denotes the 3D position of joint i of sample s, $\hat{p}^s_i$ is the corresponding recovered position. The second term $L_B = \sum_{ij} \|b^s_{ij} - \hat{b}^s_{ij}\|_2$ calculates the bone length loss, which requires the recovered bone length $\hat{b}^s_{ij}$ between joint i and j to be equal to the ground truth length b^s_{ij}. The third term $L_S = \sum_{mn,kl} \|\hat{b}^s_{mn} - \hat{b}^s_{kl}\|_2$ constrains the recovered bone lengths of the left limb must be equal to the corresponding bone lengths of recovered right limb.

The view-dependent feature and pose-dependent feature are disentangled after the SeBiReNet in the encoder. Sharing some weights before disentanglement can strengthen the interaction between h_v and h_{vi} as explained in Sec. 3.1 and make the network more compact. It's a reasonable requirement that view-dependent feature should not change the metrics of pose-dependent feature space. As the coupling operation $\otimes$ we adopt is matrix multiplication,

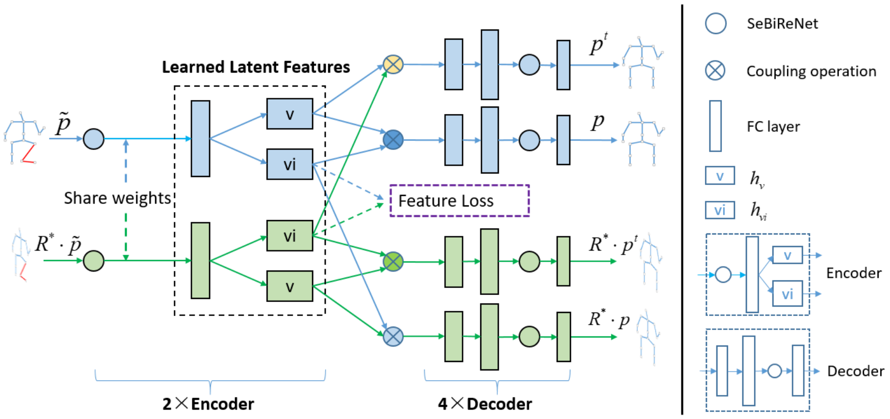

Fig. 2. The proposed architecture for the human 3D pose representation learning, which takes randomly corrupted 3D skeletons as inputs and reconstructs their correct version. Blue stream processes the non-rotated poses and green stream processes the randomly rotated poses

the requirement is satisfied only when the view-dependent feature plays a role of unit unitary transformation. For our real domain problem, we regularize the view-variant feature $h_v \in \mathbb{R}^{z \times z}$ in the $SO(z)$ space as shown in eq. 10, where $z \times z$ is the dimension of h_v. λ is a weight factor and I is an identity matrix. The orthogonal regularization is also capable of preventing the pose-related information from leaking into view-dependent feature.

$$L_O = \lambda \| I - h_v^T h_v \|_2 \tag{10}$$

To regularize the learned pose-dependent feature being view-invariant, random rotation is added to those corrupted human poses and keeping consistency between distribution $p(h_{vi}|\tilde{p})$ and $p(h_{vi}|R^* \cdot \tilde{p})$ by using a feature loss, as shown in Fig. 2. In Fig. 2, there are two pipelines to process the corrupted pose $\tilde{p}$ and the randomly rotated pose $R^* \cdot \tilde{p}$ separately. R^* is a randomly generated rotation matrix. The SeBiReNet is utilized both in the encoder and decoder. Weights are shared between all the encoders and decoders to make sure that poses under different views are encoded and decoded in the same manner. The feature loss L_f of learned pose-dependent features from different views is defined as the Frobenius norm $L_f = \|h_{vi}^1 - h_{vi}^2\|_F$. We believe that, if features are well disentangled from human pose, poses can be transfered between different views by exchanging their pose-dependent features and view-dependent features. This belief is utilized as a reinforced regularization for learning the pose representation in our method, as shown in Fig. 2 where p^t and $R^* p^t$ are view-transferred poses. Therefore, writing all together, our optimization target of learning a human 3D pose representation is formulated as eq. 11, where $L(p), L(R^* p), L(p^t), L(R^* p^t)$ are the pose reconstruction loss defined in eq. 9, $\omega_1, \omega_2, \omega_3$ are weights to adjust

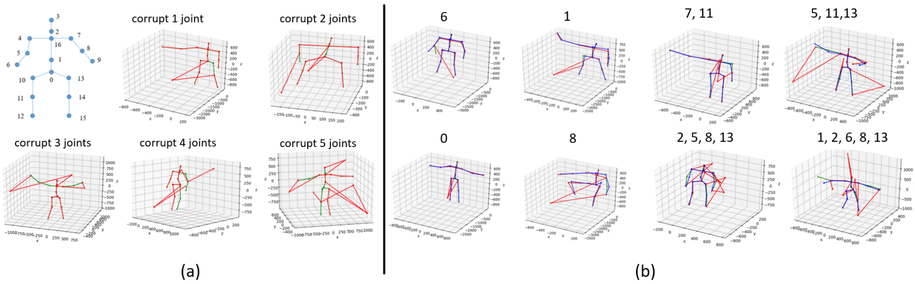

Fig. 3. (a) Illustration of the skeleton model and some generated corrupted skeleton samples, (b) Pose recovery results from randomly corrupted skeletons, the above number notes the id of corrupted joint(s). The green line, red line and blue line draw the ground truth skeleton, the corrupted skeleton and recovered skeleton, respectively. Better to view in color mode with scaling

the influence of each loss, $R(w)$ is the L2 weight regularization term to avoid overfitting.

$$\arg\min_{\theta_{vi},\theta_v,\zeta}\left\{L(p)+L(R^{*}p)+\omega_1 L(p^t)+\omega_2 L(R^{*}p^t)+\omega_3 L_f+L_O+R(w)\right\} \quad (11)$$

4 Experiments

4.1 Experimental Setup

Implementation Details. The hidden unit number of GRU cell in SeBiReNet is 32. Except the output layer, nonlinear activation function *tanh* is utilized after each MLP layer. Gradient descent optimizer with an initial learning rate of 5e-5 is used in training the DAE. Weights of different losses defined in eq. 11 are $\omega_1 = 0.01, \omega_2 = 0.01, \omega_3 = 0.1$. λ in eq. 10 is set to 0.1. The batch size is 64.
Training Set. The Cambridge-Imperial APE (Action-Pose-Estimation) dataset is used to train the proposed Siamese DAE. The Cambridge-Imperial APE dataset, which contains 245 sequences from 7 subjects performing 7 different categories of actions, is collected for 3D human pose estimation. Corrupted skeletons are generated by randomly selecting 1∼5 joints from each skeleton and moving the them to unreasonable positions with a relatively large displacement. As shown in Fig. 3 (a), these corrupted skeletons violate bio-constraints, such as bone length and allowed motion angle limits. Totally, 52500 corrupted poses are generated for training and 14000 skeletons are generated for testing. The Mean Per Joint Position Error (MPJPE) is adopted as a performance measurement of reconstructed skeletons and the trained model.
Test Sets. To verify the effectiveness of learned representations, we evaluate them on two different tasks: pose denoising and unsupervised cross-view action recognition. Two benchmark action datasets are used: Northwestern-UCLA (N-UCLA) dataset [27] and NTU RGB+D dataset [19]. Both of the two datasets

Table 1. Comparison of the performance on pose denoising among different network structures. The proposed structure achieves the best results

Network Structure	MPJPE (mm)
conventional tree (only has the recursive subnet) [28]	65.76
the diffuse subnet	64.94
concatenated structure	75.17
SeBiReNet	**42.03**
recurrent SeBiReNet	**41.58**

contain skeletons captured from different views and performed by different subjects. NTU RGB+D dataset is one of the largest skeleton datasets and N-UCLA is one of the most commonly used datasets. Pretrained encoder is applied on them to extract pose representations without any additional training, i.e., these two datasets are not used in the training phase of DAE. A 1-layer LSTM with 128 hidden units is used as the classifier in action recognition task.

4.2 Evaluation on Pose Denoising

Our model is trained and validated on the Cambridge-Imperial APE dataset. Fig. 3 (b) shows several recovered skeletons. Although we destroy the skeleton randomly and extremely, our network is still able to recover the correct positions of those invalid joints. To further show the effectiveness of our network design, we compared the performance of the proposed SeBiReNet with some baseline structures: conventional tree structure (only has the recursive part), the diffuse subnet, the concatenated structure, and the recurrent SeBiReNet. Different from the SeBiReNet which shares hidden states between the recursive subnet and the diffuse subnet, the concatenated structure takes the concatenation of the outputs from the recursive subnet and the diffuse subnet as its output. The recurrent SeBiReNet means the SeBiReNet runs in a recurrent mode as shown in Fig 1. In this experiment, we only implement it one more times.

For a fair comparison of the capability of different structures in encoding the human 3D pose, results in Table 1 is achieved by replacing the the decoder in Fig. 2 with a three-layer MLP. As shown in Table 1, even with a simple decoder, using the proposed SeBiReNet as encoder achieves an MPJPE of 42.03 mm, which is a 35% improvement compared to the first three structures in recovering corrupted skeletons. Recurrently running the SeBiReNet doesn't bring too much promotion. As skeleton data is relatively simple and low dimension, implementing the SeBiReNet only once is enough to obtain a good result. Compared to structures that only has SeBiReNet in encoder, the proposed structure in Fig. 2 which embeds the SeBiReNet both in encoder and decoder attains the best performance 33.39mm. The noteworthy result indicates that the SeBiReNet is superior to MLP layers in processing skeleton data.

To further demonstrate that the learned representation does encode the intrinsic feature of human 3D pose, we applied the pretrained network on unseen

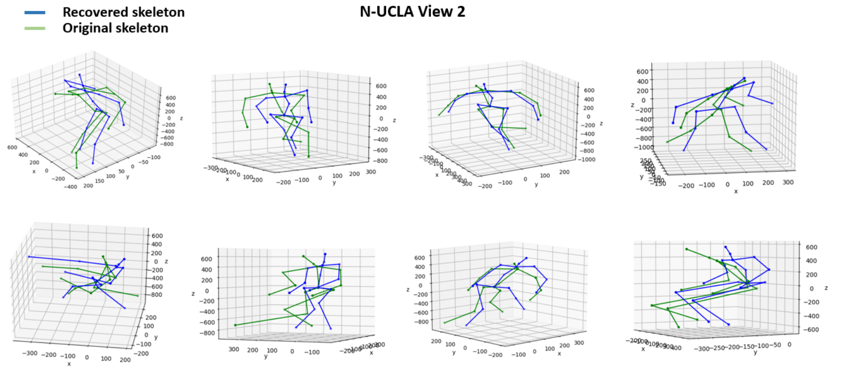

Fig. 4. Pose recovery results on N-UCLA dataset which is an unseen dataset to our pretrained DAE (Better to view in color mode with scaling)

N-UCLA dataset for a qualitative pose denoising evaluation. As Fig. 4 shows, from perspectives of fixed bone length, symmetry, and motion limit of human joints, the recovered skeletons are much more stable and reasonable compared to the original skeletons captured by the 3D sensor. The capability of denoising unseen skeleton verifies that our network has learned the intrinsic feature of human 3D pose.

4.3 Evaluation on Unsupervised Cross-View Action Recognition

To evaluate the learned pose-dependent feature, we further exploit it for unsupervised cross-view action recognition on the N-UCLA dataset and NTU RGB+D dataset. The results are shown in Table 2. In unsupervised action recognition, it's a general way to keep the pre-trained encoder fixed and only train the classifier [3,9,14]. As our target is to evaluate the performance of learned pose-dependent representation in cross-view action recognition, a simple 1-layer LSTM is adopted as classifier to reduce the influence of classifier design. Also, to this end, we only compare with those state-of-the-art methods based on RNN. Though our classifier is much simpler than those compared methods, the accuracy we achieved is competitive and even surpass some of the supervised methods. Action recognitions that are directly based on skeleton coordinates are used as baselines. Among them, the ”raw coordinates” means directly feeding the raw coordinates of skeletons into classifier. The ”normalized coordinates” means the poses are further normalized according to the mean position and standard deviation of joints. Translation of human pose is neglected when training the DAE. But for action recognition, the translation, which should be a part of the human motion, is concatenated together with the learned pose-dependent feature.

It shows explicitly in Table. 2 that the learned pose-dependent feature improves the cross-view action recognition accuracy significantly compared with baseline results, about improving by 30% on N-UCLA dataset and 10% on NTU

Table 2. Results of the cross-view action recognition on the N-UCLA dataset and the NTU RGB+D dataset (* means the result is reproduced by implementing the model reported)

Dataset		Method	Acc.(%)	# of params.
N-UCLA	Baselines	raw coordinates	38.72	-
		normalized coordinates	48.69	-
	Supervised	TLDS [4]	74.6	-
		HBRNN-L [5]	78.52	-
		Multi-task RNN [26]	87.3	-
		AGC-LSTM [20]	93.3	-
	Unsupervised	Li et al. [9]	62.5	-
		LongT GAN [35]	74.3*	-
		Denoised-LSTM [3]	76.81	-
		Ours (1-layer LSTM)	**80.30**	-
NTU RGB+D	Baseline	normalized coordinates	69.08	-
	Supervised	Hand-crafted LARP [23]	52.76	-
		LieGroups [6]	66.95	-
		Part-aware LSTM [19]	70.27	-
		ST-LSTM+TG [10]	77.70	15.37M
		Two-stream GCA-LSTM [11]	85.10	24.54M
		Bayesian GC-LSTM [34]	89.0	-
		AGC-LSTM [20]	95.0	>10.75M
	Unsupervised	LongT GAN [35]	48.1*	40.18M
		EnGAN-PoseRNN [8]	77.8	>0.7M
		Ours (1-layer LSTM)	**79.71**	0.27M

RGB+D dataset. Among those unsupervised methods on N-UCLA dataset, our method achieves the best performance with an increment of 18% compared to the work of [9]. The method of [9] is exclusively designed for learning a temporal representation using sequential skeletons in action recognition, while our method is designed for learning a representation from single pose. The accuracy of Denoised-LSTM [3] which is based on conventional DAE is quite close to our result, but the feature they learned is not view-invariant and a preprocessing treatment is needed to alleviate the influence of view changing. A similar performance is reported on NTU RGB+D dataset. Even compared with supervised methods, the accuracy is better than some of them that have more complex classifier. Performance attained on these two benchmark datasets sufficiently demonstrates the effectiveness and robustness of the learned pose representation in our method.

Though temporal information is not considered in learning pose representation, the performance in action recognition indicates that informative temporal features still can be extracted from sequential learned representations with sim-

Table 3. Ablation study based on the N-UCLA dataset. All components contribute to the overall performance and a better disentanglement

Model	Accuracy(%)	
	N-UCLA	NTU RGB+D
baseline (relative coordinates)	51.53	69.08
raw DAE	58.66	71.11
raw DAE + Feature Decomposition (FD)	60.61	73.99
raw DAE + FD + L_O	62.55	74.46
raw DAE + FD + L_O + L_f	73.81	75.72
raw DAE + FD + L_O + L_f + $L(R * p)$	76.84	77.07
full architecture	**80.30**	**79.71**

ple LSTM layer, which should be attributed to the intrinsic feature of human pose it has learned.

Moreover, as shown in Table 2, we also contrast the size of model with other state-of-the-art works that evaluated on the NTU dataset. Considering the SeBiReNet and all MLP layers used in our learning architecture, the learnable parameters in our method is about 0.27 million. As some details missed in several works, we can only estimate the lowest number of parameters in those methods, such as EnGAN-PoseRNN [8] and AGC-LSTM [20]. It can be seen that our method achieves a competitive result with the least parameters, which also shows the efficiency of our method from another perspective.

4.4 Ablation Study

To evaluate the contribution of each part in the learning architecture, we have an ablation study based on the N-UCLA and NTU RGB+D dataset as shown in Table. 3. In Table 3, the raw DAE means the structure denoted in eq. 1. "FD" means the learned latent feature is disentangled to view-dependent feature and pose-dependent feature as denoted in eq.2. L_O refers to the unit orthogonal matrix constraint on view-dependent feature, as denoted in eq. 10. L_f and $L(R* p)$ are the feature loss and reconstruction loss of randomly rotated pose as defined in Sec. 3.3. "full architecture" means integrating all the components defined in eq. 11 for a better disentanglement and representation learning.

As shown in Table 3, the raw DAE with skeleton corruption achieved an accuracy of 58.66%, which is 10% higher than the baseline result. By disentangling the latent feature and adding orthogonal loss to view-dependent feature, another 4% improvement is obtained. However, the accuracy steeply increase to 73.81% when adding the feature loss to pose-dependent feature, which indicates that the network learns better view-invariant pose feature in this case. The reconstruction losses of randomly rotated pose and generated view-transfered poses can further help improve the performance to 80.3% in cross-view action recognition, which indicates the features are better disentangled. The improvements brought by different components are steady on the NTU RGB+D dataset, but all the

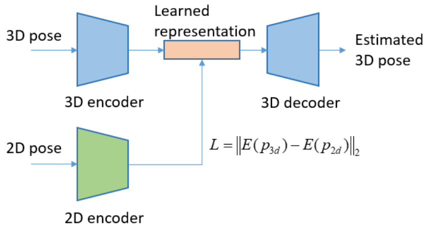

Fig. 5. Extension for 3D pose estimation

Table 4. 3D pose estimation from the generated 2D pose of H3.6M dataset

Method	MPJPE(mm)
aGCN [31]	82.9
ST-GCN [30]	57.4
Martinez et al. [15]	45.5
SemGraph [33]	43.8
Ours	53.1

components designed in our method contribute to the final performance. Feature loss and view-transferred pose losses are strong regularizations in preserving all the intrinsic pose information and learning view-invariant representations. The results, in turn, demonstrate the effectiveness of disentangling features rather than only extracting the view-invariant feature.

4.5 Extension Evaluation on 3D Pose Estimation

We further design a simple frame to explore the extension of the learned representation for 3D pose estimation from 2D pose. The extension frame contains a 3D encoder, a 2D encoder, and a decoder as shown in Fig. 5. The 3D encoder and decoder form a 3D stream and are pre-trained using 3D poses as we did in the former section. Encoder and decoder are the same with DAE in Fig. 2. In the second step, by regularizing the 2D encoder to learn a representation similar to the representation obtained in the 3D stream, 3D pose is expected to be estimated from the 2D pose. The result achieved by finetuning the 3D stream on H3.6M dataset as shown in Table 4. It can be seen that the learned representation is also applicable to the 3D pose estimation with a simple frame.

5 Conclusion

In this paper, we propose a neural network architecture to learn a human 3D pose representation by disentangling the view-dependent and pose-dependent features. Different from previous methods, the proposed method use the view-dependent and pose-dependent feature together as a pose representation for sake of preserving information. A SeBiReNet is proposed to model the human skeleton data, which considers the kinematic dependency between body joints of the human skeleton. Extensive experiments prove that the learned representation keeps the intrinsic feature of the human 3D pose and is capable of achieving excellent performance in skeleton denoising and unsupervised action recognition tasks. Utilizing the disentangled pose feature, our extension research will be focused on the view transfer between different poses.

References

1. Aberman, K., Wu, R., Lischinski, D., Chen, B., Cohen-Or, D.: Learning character-agnostic motion for motion retargeting in 2d. ACM Transactions on Graphics (TOG) **38**(4), 1–14 (2019)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013)
3. Demisse, G.G., Papadopoulos, K., Aouada, D., Ottersten, B.: Pose encoding for robust skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 188–194 (2018)
4. Ding, W., Liu, K., Belyaev, E., Cheng, F.: Tensor-based linear dynamical systems for action recognition from 3d skeletons. Pattern Recognition **77**, 75–86 (2018)
5. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1110–1118 (2015)
6. Huang, Z., Wan, C., Probst, T., Van Gool, L.: Deep learning on lie groups for skeleton-based action recognition. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6099–6108. IEEE computer Society (2017)
7. Irsoy, O., Cardie, C.: Deep recursive neural networks for compositionality in language. In: Advances in neural information processing systems. pp. 2096–2104 (2014)
8. Kundu, J.N., Gor, M., Uppala, P.K., Radhakrishnan, V.B.: Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1459–1467. IEEE (2019)
9. Li, J., Wong, Y., Zhao, Q., Kankanhalli, M.: Unsupervised learning of view-invariant action representations. In: Advances in Neural Information Processing Systems. pp. 1262–1272 (2018)
10. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal lstm with trust gates for 3d human action recognition. In: European Conference on Computer Vision. pp. 816–833. Springer (2016)
11. Liu, J., Wang, G., Duan, L.Y., Abdiyeva, K., Kot, A.C.: Skeleton-based human action recognition with global context-aware attention lstm networks. IEEE Transactions on Image Processing **27**(4), 1586–1599 (2018)
12. Liu, M., Liu, H., Chen, C.: Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognition **68**, 346–362 (2017)
13. Liu, Z., Yan, S., Luo, P., Wang, X., Tang, X.: Fashion landmark detection in the wild. In: European Conference on Computer Vision. pp. 229–245. Springer (2016)
14. Luo, Z., Peng, B., Huang, D.A., Alahi, A., Fei-Fei, L.: Unsupervised learning of long-term motion dynamics for videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2203–2212 (2017)
15. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2640–2649 (2017)
16. Nie, Q., Wang, J., Wang, X., Liu, Y.: View-invariant human action recognition based on a 3d bio-constrained skeleton model. IEEE Transactions on Image Processing (2019)

17. Ramakrishna, V., Kanade, T., Sheikh, Y.: Reconstructing 3d human pose from 2d image landmarks. In: European Conference on Computer Vision. pp. 573–586. Springer (2012)
18. Rong, Y., Liu, Z., Li, C., Cao, K., Loy, C.C.: Delving deep into hybrid annotations for 3d human recovery in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5340–5348 (2019)
19. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2016)
20. Si, C., Chen, W., Wang, W., Wang, L., Tan, T.: An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1227–1236 (2019)
21. Socher, R., Manning, C.D., Ng, A.Y.: Learning continuous phrase representations and syntactic parsing with recursive neural networks. In: Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop. vol. 2010, pp. 1–9 (2010)
22. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2602–2611 (2017)
23. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 588–595 (2014)
24. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning. pp. 1096–1103. ACM (2008)
25. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research **11**(Dec), 3371–3408 (2010)
26. Wang, H., Wang, L.: Learning content and style: Joint action recognition and person identification from human skeletons. Pattern Recognition **81**, 23–35 (2018)
27. Wang, J., Nie, X., Xia, Y., Wu, Y., Zhu, S.C.: Cross-view action modeling, learning and recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2649–2656 (2014)
28. Wei, S., Song, Y., Zhang, Y.: Human skeleton tree recurrent neural network with joint relative motion feature for skeleton based action recognition. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 91–95. IEEE (2017)
29. Xia, L., Chen, C.C., Aggarwal, J.: View invariant human action recognition using histograms of 3d joints. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on. pp. 20–27. IEEE (2012)
30. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence (2018)
31. Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph r-cnn for scene graph generation. In: Proceedings of the European conference on computer vision (ECCV). pp. 670–685 (2018)
32. Yang, X., Tian, Y.L.: Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on. pp. 14–19. IEEE (2012)

33. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3d human pose regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3425–3435 (2019)
34. Zhao, R., Wang, K., Su, H., Ji, Q.: Bayesian graph convolution lstm for skeleton based action recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6882–6892 (2019)
35. Zheng, N., Wen, J., Liu, R., Long, L., Dai, J., Gong, Z.: Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
36. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Weaklysupervised transfer for 3d human pose estimation in the wild. In: IEEE International Conference on Computer Vision, ICCV. vol. 3, p. 7 (2017)