DH-AUG: DH Forward Kinematics Model Driven Augmentation for 3D Human Pose Estimation

Linzhi Huang, Jiahao Liang, and Weihong Deng*

Beijing University of Posts and Telecommunications
{huanglinzhi, jiahao.liang, whdeng}@bupt.edu.cn

Abstract. Due to the lack of diversity of datasets, the generalization ability of the pose estimator is poor. To solve this problem, we propose a pose augmentation solution via DH forward kinematics model, which we call DH-AUG. We observe that the previous work is all based on single-frame pose augmentation, if it is directly applied to video pose estimator, there will be several previously ignored problems: (i) angle ambiguity in bone rotation (multiple solutions); (ii) the generated skeleton video lacks movement continuity. To solve these problems, we propose a special generator based on DH forward kinematics model, which is called DH-generator. Extensive experiments demonstrate that DH-AUG can greatly increase the generalization ability of the video pose estimator. In addition, when applied to a single-frame 3D pose estimator, our method outperforms the previous best pose augmentation method. The source code has been released at https://github.com/hlz0606/DH-AUG-DH-Forward-Kinematics-Model-Driven-Augmentation-for-3D-Human-Pose-Estimation.

Keywords: Pose Augmentation, Video, Forward Kinematics, Human Pose Estimation

1 Introduction

3D pose estimation is the task of estimating 3D human pose from images. It is a fundamental task in action recognition [22,50,38,19], human tracking [30], etc. It is difficult to obtain a 3D label, so the existing 3D data is very limited and the diversity is seriously insufficient. This also leads to poor generalization ability of the 2D-to-3D model.

Recently, a work [23] enhanced data by randomly exchanging limbs, locally rotating limbs, and randomly changing bone length. This method is dependent on the random seed, and the result is unstable. PoseAug [11] uses GAN [12] to solve the above problems. However, PoseAug is also designed for a single-frame 3D pose estimator. There are some problems that can not be ignored in pose augmentation in video 3D human pose estimation: angle ambiguity and angle continuity. Most pose discriminators [5,47,11] calculates the cosine angle value through the inner product of two bone vectors for constraint. But this

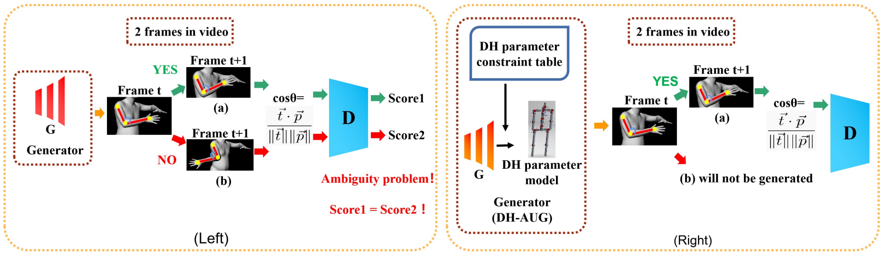


Fig. 1. Angle ambiguity (multiple solutions). **Left**: Pose augmentation of ordinary GAN framework. **Right**: DH-AUG. **(a)**: Elbow rotates normally (angle is about 90°). **(b)**: Elbow rotates abnormally (angle is about -90°). Although the rotation directions of (a) and (b) are different, the cosine angle values calculated by vector inner product are the same. Both of them will make the discriminator output the same score, resulting in ambiguity. To weaken the ambiguity problem, we improve the generator by adding DH forward kinematics model (DH parameter model) and constraints. t and p are a pair of adjacent bone vectors. SMPL [25] is only used for visualization.

is a problem of multiple solutions (angle ambiguity) as shown in Fig. 1 (Left). The value calculated by the inner product corresponds to multiple angles. For example, 0 corresponds to 90° and -90°, which makes the discriminator unable to distinguish between elbow 90° internal rotation and 90° external rotation. Both of them will make the discriminator output the same score, and the data distribution of the generator will contain the angle value of abnormal rotation. What's more, there will be discontinuous actions in the skeleton video because of angle ambiguity. So it is not enough to use the discriminator for constraints. We try to modify the generator to weaken this problem. Specifically, we use DH parameters to build a human kinematics model (DH parameter model). This model allows us to obtain a new pose directly by changing the joint angle, and we can easily constrain the rotation direction of the joint. We introduce this model into the generator and constrain the DH parameters so that the generator will not produce an unreasonable pose as shown in Fig. 1 (Right). Inspired by some previous work [5,39,45], we also add timing information to the discriminator to increase the continuity of the generated skeleton video.

Our contributions are as follows:

- We propose DH-AUG: a pose augmentation framework for 3D human pose estimation. It consists of DH-Generator, DH parameter model, single-frame pose discriminator and multi-stream motion discriminator.
- We use DH parameters to design a human kinematics model, called DH parameter model. By adding DH parameter model and constraints to the generator, the angle ambiguity is successfully weakened and the possibility of generating unreasonable pose is reduced.
- Extensive experiments demonstrate that DH-AUG can greatly increase the generalization ability of the video pose estimator. In addition, when applied

to a single-frame 3D pose estimator, our method outperforms the previous best pose augmentation method.
- We release a new dataset (DH-3DP) synthesized with DH-AUG, which can be used in the 2D-to-3D network.

2 Related Work

3D human pose estimation. There are two mainstream monocular 3D human pose estimation methods, one is to obtain 3D pose end-to-end [33,43,44], and the other is through the multi-stage method, first obtain 2D pose from the images [40,6,42], and then further obtain 3D pose from 2D pose [28,20,48,5]. The second method is more common. We do not pay too much attention to the model structure. We focus on pose augmentation for 2D-to-3D networks and produce 2D-3D pairs. According to the input mode, it can be divided into single-frame input and video input. Video input can weaken the depth ambiguity problem [34,52,2]. We design a pose augmentation scheme for single-frame pose estimation and video pose estimation.

Kinematic model. The kinematic model is widely used in the field of the robot [9], hand pose estimation [32,18], and games. Recent work [21] uses forward and inverse kinematics to make up for the shortcomings of 3D pose estimation and mesh parameter models. Inspired by this, we use the DH parameter [8] to build a 3D human forward kinematics model to weaken the angle ambiguity. DH parameter is a method to describe the coordinate system of connecting links.

pose augmentation for 3D human pose estimation. Due to the high cost of 3D data acquisition and insufficient data diversity, the 2D-to-3D model is difficult to have good generalization ability. In some works, pose augmentation of 3D pose estimation is carried out by synthesizing images [35,36,46]. It is worth noting that there is another way to obtain new data pairs by synthesizing 2D and 3D data. The recently proposed evolutionary algorithm [23] uses random exchange, local rotation to generate data. The data generated in this way has great randomness, depending on the preset parameters. PoseAug[11] proposes to use GAN with a feedback mechanism to generate data, which is more effective than the former. However, this method has insufficient constraints on joint rotation. This is not conducive to being extended to video pose estimation. Therefore, we propose a combination of the DH parameter model and GAN for pose augmentation.

3 Method

3.1 Overview

There are multiple solutions for mapping the coordinates of 3D keypoints to the angle value, so it is not enough to use the discriminator for constraints. To weaken the angle ambiguity problem and further improve the effect of pose augmentation, we introduce DH parameters into GAN framework, as shown in

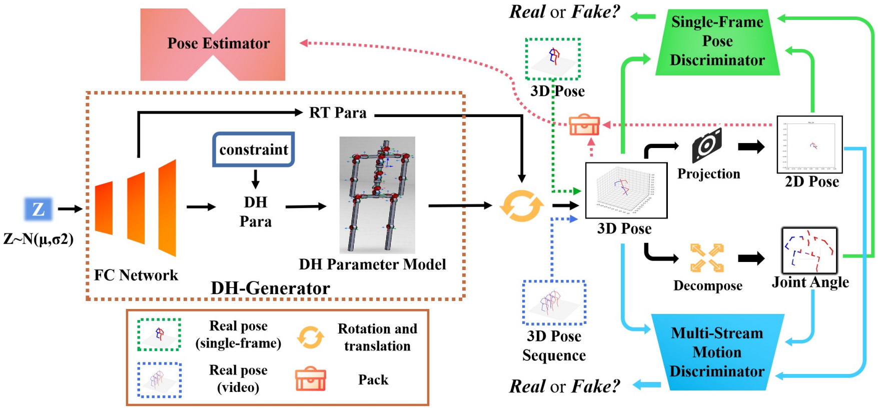
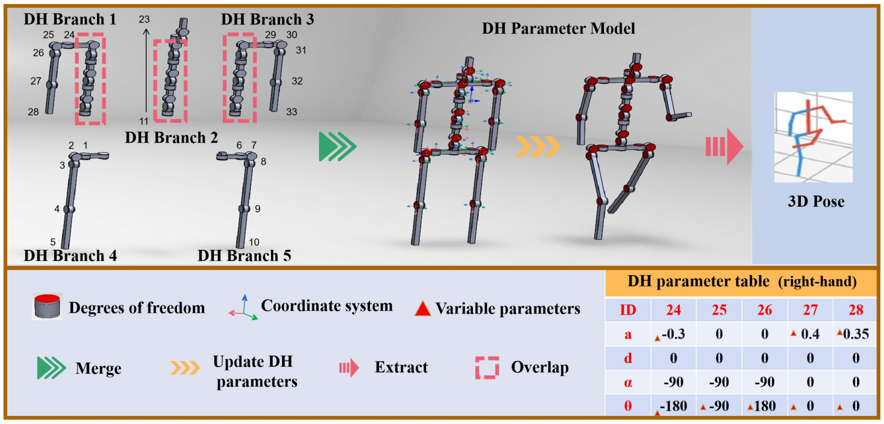


Fig. 2. Overview of the overall framework of DH-AUG. 128-dimensional vectors are sampled from the normal distribution and input into the fully connected network to obtain DH parameters, global rotation and translation parameters. Then, the 3D pose is obtained through DH parameter model. **1) Single-frame**: The 3D pose, 2D pose and joint angle are transmitted to the single-frame pose discriminator for training. **2) Video**: Input 3D pose sequence, 2D pose sequence, bone rotation trajectory (joint angle) into single-frame pose discriminator and multi-stream motion discriminator. Finally, the newly generated 2D-3D data pair is packaged into a new dataset and transmitted to the pose estimator for training.

Fig. 2. We use the fully connected network to generate DH parameters, etc., and transfer them into the DH parameter model to obtain the corresponding 3D pose. In addition, we also use discriminators to force the generator to generate more reasonable and diversified 3D pose. It is worth noting that we add constraints to the DH parameter model to avoid generating unreasonable pose and weaken the angle ambiguity. More specific contents will be introduced in this section.

3.2 DH Parameter Model

Human kinematics model based on DH parameters. DH parameter [8] is a method to describe the coordinate system of connecting links. The schematic diagram of the DH parameter can be seen in the right part of Fig. 4, where a is the link length, d is the link offset, α is the twist angle, θ is the joint angle. These four parameters are DH parameters. Each degree of freedom (DOF) has a set of DH parameters. We use DH parameters to establish the human kinematics model as shown in Fig. 3. Some parameters of the model are fixed, which determines the connection relationship between bones, while others determines the rotation relationship between bones and the length of bones. In the DH parameter table in Fig. 3, those marked with red triangles are variable parameters, and others are preset fixed parameters. See Alg. 1 for the process of building a human kinematics model (DH parameter model) with DH parameters. Δa, Δd, $\Delta \alpha$, $\Delta \theta$



DH parameter table (right-hand)

ID	24	25	26	27	28
a	▲-0.3	0	0	▲ 0.4	▲0.35
d	0	0	0	0	0
α	-90	-90	-90	0	0
θ	▲-180	▲ -90	▲180	▲ 0	▲ 0

Fig. 3. DH parameter model. There are 33 degrees of freedom (DOF) and 48 changeable DH parameters. 5 DOF in the figure are not drawn (head, ankle, and wrist). We built 5 DH branches. The root node is the hip, and the overlapping parts share DH parameters. The part sharing DH parameters combines 5 branches into a complete human kinematics model. Then the new transformation matrix is obtained by updating the DH parameters. Finally, a new 3D pose is extracted from the transformation matrix. (See Alg. 1 for the process of building a human kinematics model with DH parameters. The complete DH parameter table is in the supplementary material.)

are the change in DH parameters. R_x, R_y, R_z are the global rotation parameters. T_x, T_y, T_z are the global translation parameters. In addition, they are the values output by the fully connected network. Output P_{new} is a new 3D pose. The value of N_{branch} is 5, which is the number of DH branches. $N_{Dof(i)}$ is the number of degrees of freedom (DOF) per branch. First, the DH parameters are converted into the transformation matrix:

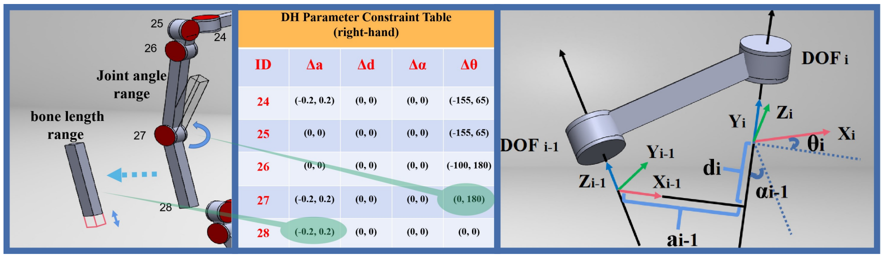
$$M_{DH} = \begin{bmatrix} cos(\theta) & -sin(\theta) & 0 & a \\ sin(\theta)cos(\alpha) & cos(\theta)cos(\alpha) & -sin(\alpha) & -dsin(\alpha) \\ sin(\theta)sin(\alpha) & cos(\theta)sin(\alpha) & cos(\alpha) & dcos(\alpha) \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

where a is the link length, d is the link offset, α is the twist angle, θ is the joint angle. Next, The inner product is used to update the transformation matrix:

$$M'_{DH(i,k+1)} = M_{DH(i,k)} M_{DH(i,k+1)} \tag{2}$$

where i is the index of the branch, and k is the index of the degree of freedom in $branch_i$. Then, a new 3D pose is extracted from M_{DH}. Finally, we globally rotate and translate the new 3D pose. Other details are illustrated in Fig. 3.

Constraints on DH parameter model. We implemented two constraints on the DH parameter model. **1)** We removed the redundant degrees of freedom (DOF). For example, we only set 1 DOF for the elbow and knee instead of 3, and



DH Parameter Constraint Table (right-hand)

ID	Δa	Δd	Δα	Δθ
24	(-0.2, 0.2)	(0, 0)	(0, 0)	(-155, 65)
25	(0, 0)	(0, 0)	(0, 0)	(-155, 65)
26	(0, 0)	(0, 0)	(0, 0)	(-100, 180)
27	(-0.2, 0.2)	(0, 0)	(0, 0)	(0, 180)
28	(-0.2, 0.2)	(0, 0)	(0, 0)	(0, 0)

Fig. 4. Left: The constraint diagram of the elbow. **Middle**: DH parameter constraint table. **Right**: The schematic diagram of DH parameter [8]. a is link length, d is link offset, α is twist angle, θ is the joint angle. $\Delta a, \Delta d, \Delta\alpha, \Delta\theta$ is the change in the DH parameter. **(The complete DH parameter constraint table can be seen in the supplementary material.)**

```
Algorithm 1 DH parameter model
Input:$\Delta a, \Delta d, \Delta\alpha, \Delta\theta, R_x, R_y, R_z, T_x, T_y, T_z$
Output:$P_{new}$
  for $i$ in $N_{branch}$ do
    for $k$ in $N_{Dof(i)}$ do
      $A = a_{i,k}+\Delta a_{i,k}$; $B = d_{i,k}+\Delta d_{i,k}$; $C = \alpha_{i,k}+\Delta\alpha_{i,k}$; $D = \theta_{i,k}+\Delta\theta_{i,k}$;
      $M_{DH(i,k)}$ = Get_Matrix($A$, $B$, $C$, $D$); See Eq.1
    for $k$ in $N_{Dof(i)}$ - 1 do
      $M_{DH(i,k+1)}$ = Update_MDH($M_{DH(i,k)}$, $M_{DH(i,k+1)}$); See Eq.2
  for $i$ in $N_{branch}$ do
    for $k$ in $N_{Dof(i)}$ do
      $x_{i,k} = M_{DH(i,k,0,3)}$; $y_{i,k} = M_{DH(i,k,1,3)}$; $z_{i,k} = M_{DH(i,k,2,3)}$;
      $P_{new(i,k)} = R_x\ R_y\ R_z(x_{i,k},\ y_{i,k},\ z_{i,k}) + (T_x,\ T_y,\ T_z)$;
  return $P_{new}$
```

the number of DOF is changed from 48 to 33 (the number of key points is 16). For details, see the human skeleton in Fig. 3. This operation not only greatly reduces the parameters that the GAN needs to learn, but also prevents the generator from producing a human skeleton with unreasonable rotation direction. **2)** We designed a DH parameter constraint table to limit the value of DH parameters. We list the constraint table of the right-hand branch, as shown in Fig. 4. The left side of Fig. 4 is the constraint diagram of the elbow, and the middle side is the DH parameter constraint table of the whole right-hand branch. DH parameter constraint table is added to the last layer of fully connected network:

$$P_{DH} = (1 + \tanh(O_{FC}))\frac{T_{DH(max)} - T_{DH(min)}}{2} \tag{3}$$

where P_{DH} is the DH parameter, O_{FC} is the output of the fully connected network, T_{DH} is the DH parameter constraint table (Fig. 4).

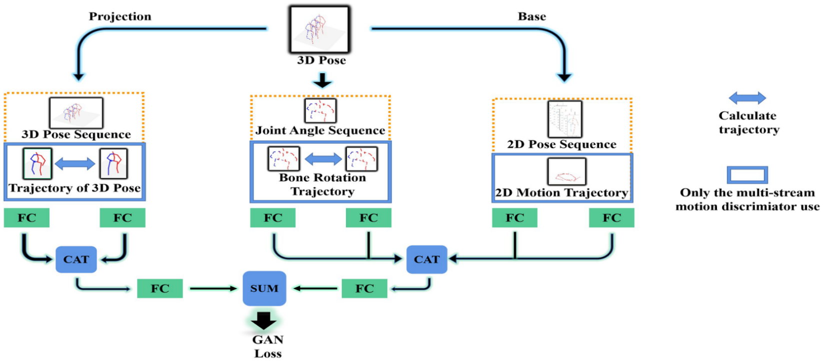


Fig. 5. Multi-stream motion discriminator (MSMD). It has 3 two-stream branches. **Left**: 3D pose two-stream branch. **Middle**: Bone rotation two-stream branch. **Right**: 2D pose two-stream branch. When we remove the branch in the blue box, it becomes a single-frame pose discriminator.

3.3 Architecture

Fig. 2 shows the overall framework of DH-AUG. Here we will describe various components of DH-AUG.

DH-generator. We combine the DH parameter model and fully connected network (FC) to form a new generator called DH-generator. The fully connected network first samples the 128-dimensional vector z from the normal distribution as the input then generates DH parameters: Δa, Δd, $\Delta \alpha$, $\Delta \theta$ (48 in total, includes bone length), global rotation parameters: R_x, R_y, R_z, and global translation parameters: T_x, T_y, T_z. These parameters are input into the DH parameter model. After a series of operations described in Section 3.2, a new 3D pose will be generated. Finally, the 3D pose is projected [14] to a 2D pose through camera parameters (from the original dataset). When DH-generator is used in single-frame pose estimation, only one 3D pose is generated by sampling one vector. However, when DH-generator is used in video pose estimation, a vector is sampled to generate a 3D pose sequence.

Multi-stream motion discriminator (MSMD). To generate a skeleton video with motion continuity, we add timing information to the discriminator, as shown in Fig. 5). It has 3 two-stream branches. 1) 3D pose two-stream branch. We input the 3D pose sequence and the trajectory of the 3D pose into this two-stream branch. The trajectory of the 3D pose is calculated as follows:

$$D_{3D} = \sum_{t=1}^{T} \sum_{i=0}^{I} (P_{3D_{(t,i)}} - P_{3D_{(t-1,i)}}) \tag{4}$$

where $P_{3D_{(t,i)}}$ is the 3D coordinate of the ith key point in frame t. 2) Bone rotation two-stream branch. We calculate the joint angle between adjacent bones, and the formula is as follows:

$$A_{(t,i)} = \frac{V_{t,i} \cdot V_{t,i-1}}{L_{t,i} L_{t,i-1}} \tag{5}$$

where $V_{t,i}$ is the ith bone vector in frame t, $L_{t,i}$ is the ith bone length in frame t. i and $i-1$ are a pair of adjacent bones. Another input to this two-stream branch is the bone rotation trajectory:

$$D_{Angle} = \sum_{t=1}^{T} \sum_{i=0}^{I} (A_{t,i} - A_{t-1,i}) \tag{6}$$

where $A_{t,i}$ is the ith joint angle in frame t. 3) 2D pose two-stream branch. We input the 2D pose sequence and 2D motion trajectory into this two-stream branch. This branch mainly guides the generator to produce the correct viewpoint. The calculation formula of 2D motion trajectory is as follows:

$$D_{2D} = \sum_{t=1}^{T} (P_{2D_{(t,root)}} - P_{2D_{(t-1,root)}}) \tag{7}$$

where $P_{2D_{(t,root)}}$ is the 2D coordinate of the root key point in frame t, *root* represents the key point of the hip.

Single-frame pose discriminator. The single-frame pose discriminator we use is a simplified version of the MSMD. Its structure is the content after removing the components in the blue box in Fig. 5.

Training loss. Loss used by our GAN is the objective function in improved Wasserstein GAN [13]. The loss we finally use is as follows:

$$\gamma = \begin{cases} 1 & epoch >= \beta \\ 0 & epoch < \beta \end{cases} \tag{8}$$

$$\begin{aligned} L = {} & E[D_s(X_f)] - E[D_s(X_r)] + \alpha E[(\left\|\nabla_{\hat{X}} D_s(\hat{X})\right\|_2 - 1)^2] \\ & + \gamma(E[D_m(X_f)] - E[D_m(X_r)] + \alpha E[(\left\|\nabla_{\hat{X}} D_m(\hat{X})\right\|_2 - 1)^2]) \end{aligned} \tag{9}$$

where D_s represents the output of single-frame pose discriminator, D_m represents the output of multi-stream motion discriminator, α represents the weight of gradient penalty, γ represents whether to turn on the multi-stream motion discriminator, X_f is fake data, X_r is real data, $\hat{X}$ is randomly sampled data, β is the epoch that turns on the multi-stream motion discriminator. In our experiment, α is 10, β is 4.

Pose estimator. In this paper, we use SemGCN [51], SimpleBaseline [28] and VPose [34] as single-frame 3D pose estimators, VPose [34] and PoseFormer [52] as video 3D pose estimators, and Det [10], CPN [4], HR [41] and ground truth as 2D pose estimators.

About the use of synthetic data. Each epoch generates the same number of data pairs as the training set and packs them into a new dataset. Then in the next epoch, we will train the 3D pose estimator on the new dataset and the original dataset.

4 Experiments

4.1 Implementation Details

We use the fully connected network as the backbone network. See supplementary material for the specific structure of generator and discriminator. When pose augmentation is performed for the video pose estimator, we first train the single-frame pose discriminator for 4 epochs and then turn on the multi-stream motion discriminator. Single-frame: batch size is 1024, video: batch size is 512. The pose estimator uses the Adam optimizer with a learning rate of 1e-4, 1e-3, or 2e-3. The first 50 epochs use linear attenuation, and the subsequent epochs attenuate each epoch by 5% to 10%. Both generator and discriminator use Adam optimizer, and the learning rate remains 1e-4 unchanged. The training is carried out on one 1080ti GPU. Training about 100 to 140 epochs. The data used to train 2D-3D pose lifting network and DH-AUG are consistent. For example, In the weakly-supervised settings, both the pose lifting network and DH-AUG are trained using S1 in H36M. See supplementary material for DH parameter constraint table, model structure, etc.

4.2 Datasets

Human3.6M [15] is the largest benchmark dataset. Subjects 1, 5, 6, 7, 8 are used as the training set, and subjects 9, 11 are used as the test set. In case of weak supervision, S1 or S1, S5 shall be used for training, and S9, S11 shall be used for evaluation. MPJPE was used as evaluation criteria.

MPI-INF-3DHP [30] and **3DPW** [27] are large 3D datasets containing complex outdoor scenes. Instead of using them for training, we use their test sets to evaluate the model's generalization ability to unseen environments. Evaluation criteria: PCK, AUC, MPJPE (MPI) and PA-MPJPE (3DPW).

LSP [16] and **MPII** [1] are two 2D pose datasets containing a large number of outdoor scenes. We selected several difficult pictures for qualitative experiments.

DH-3DP: We synthesized a dataset with more than 1 million 2D-3D data pairs using DH-AUG. The synthesis method of this dataset is: S15678 of H36M is used as the training set to train DH-AUG, with a total of 110 epochs. We use the pretrained DH-AUG to generate more than 1 million 2D-3D data pairs. See the supplementary materials for more details.

4.3 Pose Augmentation in Video Pose Estimation

We use VPose [34] and PoseFormer [52] as the 3D pose estimators and Det [10], CPN [4], HR [41], and ground truth as the 2D pose estimators. Experiments were carried out with 9 and 27 frames. Because H36M is large, we choose to use 10 times of downsampling data for training. The results are shown in Table. 1. It can be seen that DH-AUG can greatly increase the generalization ability of the video pose estimator. It is worth noting that PoseAug [11] is designed for a single-frame pose estimator. It can not be directly used in a video pose

Table 1. Results of using DH-AUG in video 3D pose estimation. f represents the number of input frames. The evaluation criteria uses MPJPE. We downsample the frames used by a factor of 10. We use VPose [34] and Poseformer [52] as 3D pose estimators. And DET [10], CPN [4], HR [41] and GT are used as 2D pose estimators. (It is worth noting that PoseAug [11] is designed for single-frame pose estimator.)

	MPI-3DHP (↓)				H36M (↓)			
Method	DET	CPN	HR	GT	DET	CPN	HR	GT
VPose [34] (f=9)	97.56	94.26	90.83	90.7	61.47	55.74	53.79	42.14
Vpose+DH-AUG (f=9)	**84.23**	**84.76**	**82.57**	**80.39**	**60.81**	**55.66**	**53.04**	**41.21**
VPose [34] (f=27)	101.99	97.33	94.62	91.76	61.84	56.57	52.89	42.18
Vpose + DH-AUG (f=27)	**86.34**	**88.38**	**84.37**	**80.85**	**61.19**	**56.07**	**52.57**	**41.52**
PoseFormer [52] (f=9)	95.09	88.01	82.38	85.28	63.28	56.47	54.24	42.02
PoseFormer + DH-AUG (f=9)	**81.99**	**81.13**	**76.07**	**76.25**	**63.13**	**55.73**	**53.32**	**39.29**
PoseFormer [52] (f=27)	92.71	86.38	83.16	84.67	62.26	55.00	53.34	39.63
PoseFormer + DH-AUG (f=27)	**81.04**	**77.13**	**72.18**	**75.36**	**62.26**	**54.95**	**52.46**	**37.92**

Table 2. Results on H36M and MPI. Evaluation criteria: MPJPE. Best in bold.

Method	3DHP (↓)	H36M (↓)
VPose [34] (f=27)	91.76	42.18
Liu et al [24] (f=243)	91.86	42.70
Anatomy [3] (f=27)	86.01	39.98
PoseFormer [52] (f=27)	84.67	39.63
PoseFormer (f=27) + **DH-AUG (Ours)**	**75.36**	**37.92**

estimator. Table. 2 is the result on H36M and MPI. It can be seen that our method outperforms other SOTA methods.

4.4 Pose Augmentation in Single-Frame Pose Estimation

To be consistent with other methods, we use HR [41] as the 2D pose estimator and VPose [34] as the 3D pose estimator. Table. 3 is the result on H36M. It can be seen that our method outperforms other SOTA methods (fully-supervised).

To evaluate the model's generalization ability, we only use H36M for training and use MPI and 3DPW as test sets. Moreover, we use ground truth as 2D data and VPose [34] as the 3D pose estimator. See Table. 4 for MPI test results. See

Table 3. Results on H36M (fully supervised). Evaluation criteria: MPJPE. Best in bold. * denotes the SOTA pose augmentation method.

Method	MPJPE (↓)
SemGCN (CVPR'19) [51]	57.60
Sharma (CVPR'19)[37]	58.00
Moon (ICCV'19) [31]	54.40
VPose (CVPR'19) [34]	52.70
*Li (CVPR'20) [23]	50.90
*VPose + PoseAug (CVPR'21) [11]	50.20
VPose + DH-AUG	**49.81**

Table 4. Results on MPI (fully supervised). The evaluation criteria were PCK, AUC and MPJPE. CE means evaluation across datasets. Best in bold. * represents SOTA pose augmentation method. **S1 + S5**: Use S1 and S5 for training.

Method	CE	MPJPE (↓)	PCK (↑)	AUC (↑)
Mehta[29]		117.60	76.50	40.80
VNect [30]		124.70	76.60	40.40
Multi Person [6]		122.20	75.20	37.80
OriNet [26]		89.40	81.80	45.20
LCN [7]	✓	-	74.00	36.70
HMR [17]	✓	113.20	77.10	40.70
SRNet [49]	✓	-	77.60	43.80
RepNet [47]	✓	92.50	81.80	54.80
*Li [23]	✓	99.70	81.20	46.10
VPose [34]	✓	86.60	-	-
*VPose+PoseAug [11]	✓	73.00	88.60	57.30
VPose+DH-AUG (S1+S5)	✓	72.93	88.60	57.65
VPose+DH-AUG	✓	**71.17**	**89.45**	**57.93**

Table 5. Results on H36M and MPI (weakly supervised). Evaluation criteria:MPJPE. Best in bold.

Train Set	S1		S1 + S5	
Method	MPI	H36M	MPI	H36M
VPose [34]	116.4	65.2	93.5	57.9
VPose+PoseAug [11]	90.3	56.7	77.9	51.3
VPose+DH-AUG	**86.72**	**52.15**	**72.93**	**46.99**

the right part of Table. 6 for 3DPW test results. We can observe that our method achieves the best performance under all the metrics.

The effect of our method is more obvious when it is weakly-supervised. Consistent with other pose augmentation methods, we used S1 or S1, S5 in H36M dataset for training and evaluated on H36M and MPI. In addition, we use ground truth as 2D data and VPose [52] as the 3D pose estimator. The results are shown in Table. 5. It can be seen that DH-AUG outperforms the previous best method.

To further prove the generality of our method. We use SemGCN [51], SimpleBaseline [28] and, VPose [34] as the 3D pose estimator and Det [10], CPN [4], HR [41], and ground truth as the 2D pose estimator. The results are shown in Table. 6. It can be seen that our method outperforms the previous best pose augmentation method.

Analysis of the data distribution. To further verify the diversity of the data we generated, we visualized the data distribution of the left knee and right knee. Distribution of H36M (before augmentation) form a small and concentrated cluster, also showing a limited diversity (Fig. 6 (b)). However, our method (DH-AUG) obtains a huge and decentralized cluster as shown in Fig. 6 (d). This shows that DH-AUG generates more diverse pose, and also proves why our method can greatly enhance the generalization ability. The comparison of distribution before and after adding constraints will be introduced in section 4.6. In addition, we provide the distribution of all joint angles, as shown in Fig. 6 (e) and (f).

Table 6. Results on H36M, MPI and 3DPW. Different 2D and 3D pose estimators were used to evaluate the results before and after using DH-AUG. Consistent with previous experiments, DET [10], CPN [4], HR [41] and GT are used as 2D pose estimators, SemGCN [51], SimpleBaseline [28] and VideoPose [34] are used as 3D pose estimators, and +PoseAug denotes the result of the recent SOTA pose augmentation method [11]. The evaluation criteria is MPJPE (MPI, H36M) and PA-MPJPE (3DPW).

	MPI-3DHP (↓)				H36M (↓)				3DPW (↓)
Method	DET	CPN	HR	GT	DET	CPN	HR	GT	GT
SemGCN [51]	101.90	98.70	95.60	97.40	67.50	64.70	57.50	44.40	102.00
SemGCN + PoseAug [11]	89.90	89.30	89.10	86.10	65.20	60.00	55.00	41.50	82.20
SemGCN + DH-AUG	**79.68**	**76.67**	**72.99**	**71.31**	**63.16**	**56.93**	**54.04**	**40.00**	**79.07**
SimpleBaseline [28]	91.10	88.80	86.40	85.30	60.50	55.60	53.00	43.30	89.40
SimpleBaseline + PoseAug [11]	78.70	78.70	76.40	76.20	58.00	53.40	51.30	39.40	**78.10**
SimpleBaseline + DH-AUG	**77.99**	**75.87**	**72.97**	**72.28**	**57.86**	**53.13**	**50.06**	**38.89**	80.52
VPose [34] (1-frame)	92.60	89.80	85.60	86.60	60.00	55.20	52.70	41.80	94.60
VPose + PoseAug [11]	78.30	78.40	73.20	73.00	57.80	52.90	50.20	38.20	81.60
VPose + DH-AUG (1-frame)	**76.70**	**74.82**	**71.07**	**71.17**	**57.66**	**52.52**	**49.81**	**37.01**	**79.28**

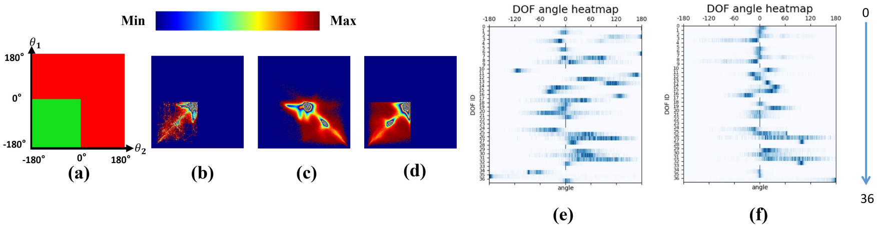


Fig. 6. Data distribution. (b), (c), (d) is the data distribution of left knee and right knee joint angles. The normal rotation range of the knee is -180° to 0°. (e), (f) is the data distribution of all joint angles. The amount of data in the (b), (c), (d) is the same. **(a)**: The green area is the normal area. The red area is the area where ambiguity occurs. $\boldsymbol{\theta_1}$: Left knee joint angle. $\boldsymbol{\theta_2}$: Right knee joint angle. **(b)**: Data distribution of H36M datasets (before pose augmentation). **(c)**: Pose augmentation (no constraints). **(d)**: Pose augmentation (add constraints). **(e)**: Pose augmentation (no constraints). **(f)**: Pose augmentation (add constraints).

4.5 Qualitative Results

We select difficult figures from several datasets for estimation, as shown in Fig. 7. The pose estimator enhanced with DH-AUG can get results with more correct action, better scale matching, and higher accuracy. **More qualitative results are shown in the supplementary material.**

We selected 2 frames of data close to the camera and found that DH-AUG can solve the scale problem shown in Fig. 8. The reason for the scale mismatch is that the human motion in H36M is concentrated in one range, which makes the model unable to fully learn the relationship between bone length and distance.

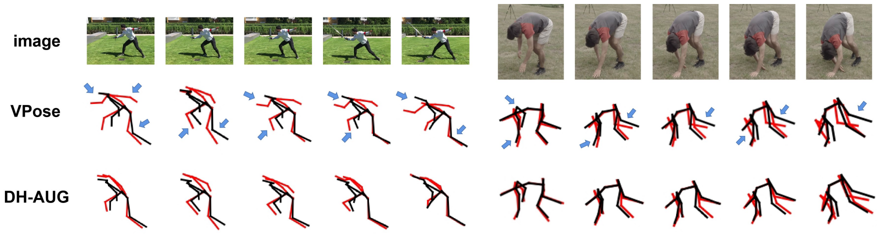


Fig. 7. Qualitative results on MPI and 3DPW. The black pose is the ground truth. The blue arrow points to the location of the main correction. **More qualitative results are shown in the supplementary material.**

Table 7. Ablation study. DHG: DH-generator. **BR**: Bone rotate two-stream module. **2DP**: 2D Pose two-stream. **DHT**: DH parameter constraint table.

Method	3DHP (↓)	H36M (↓)
Baseline	90.70	42.14
+ DHG + BR	88.97	43.65
+ DHG + BR + 2DP	84.12	41.31
+ DHG + BR + 2DP + 3DP	82.86	41.20
+ DHG + BR + 2DP + 3DP + DHT	80.39	41.21

4.6 Ablation Study

BR, 2DP, 3DP. BR is the bone rotation two-stream branch, which is used to constrain the joint angle parameters produced by the generator. **2DP** is the 2D pose two-stream branch, which is mainly used to constrain global translation parameters and motion trajectories. **3DP** is the 3D pose two-stream branch, which is mainly used to constrain global rotation parameters and enable the model to learn bone length information.

Effect of DH parameter constraint table. DHT is the DH parameter constraint table. Fig. 6 is the data distribution of the left knee and right knee. The normal rotation range is about -180° to 0°. Before adding the DHT (Fig. 6 (c)), the data distribution is asymmetric and unreasonable. Fig. 6 (c) has a lot of outward rotation values (between 0° and 180°, the knee cannot be external rotation), which indicates that the generator has learned the wrong human kinematics information. This causes the generator to produce the skeleton video shown in Fig. 9 (a). However, by observing Fig. 6 (d), it will be found that the distribution is symmetrical and reasonable. After adding constraints, the skeleton video generated by DH-AUG is shown in Fig. 9 (b).

By observing the table. 7, we can see that the performance is improved by gradually adding **BR**, **2DP**, **3DP**, and **DHT** modules, which verifies the effectiveness of these modules.

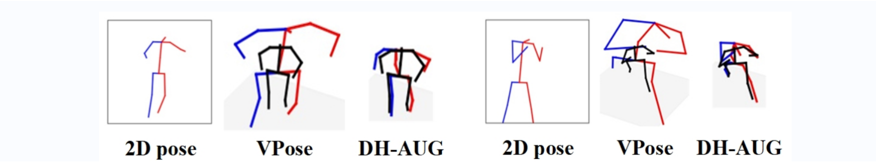


Fig. 8. Scale problem. Columns 1, 4 are 2D poses, columns 2, 5 are the results before pose augmentation, and columns 3, 6 are the results of using DH-AUG. The black pose is the ground truth.

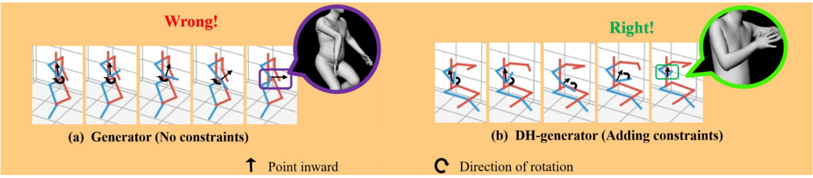

Wrong!
Right!
(a) Generator (No constraints)
(b) DH-generator (Adding constraints)
Point inward
Direction of rotation

Fig. 9. Skeleton video generated by DH-AUG. (a) No constraints. **(b)** Adding constraints. SMPL is only used for visualization. **See the supplementary material for more skeleton videos.**

4.7 Limitation Analysis

Although our method increases the generalization ability of 3D human pose estimation, our method still has some limitations. The DH parameter constraint table we use is manually set according to personal experience. This increases the number of hyper-parameters that need to be adjusted. Although it will not have a great impact on the final result, it increases some workload.

5 Conclusion

In this paper, we propose a pose augmentation solution, which we call DH-AUG. DH-AUG has a special kinematics model called the DH parameter model, which weakens the angle ambiguity (multiple solutions). We use 3 common single-frame 3D pose estimators and 2 video 3D pose estimators to experiment. Extensive experiments demonstrate that DH-AUG can greatly increase the generalization ability of the pose estimator.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 62192784 and Grant 61871052.

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. pp. 3686–3693 (2014)
2. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.J., Yuan, J., Thalmann, N.M.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2272–2281 (2019)
3. Chen, T., Fang, C., Shen, X., Zhu, Y., Chen, Z., Luo, J.: Anatomy-aware 3d human pose estimation with bone-based pose decomposition. IEEE Transactions on Circuits and Systems for Video Technology **32**(1), 198–209 (2021)
4. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7103–7112 (2018)
5. Cheng, Y., Yang, B., Wang, B., Tan, R.T.: 3d human pose estimation using spatio-temporal networks with explicit occlusion training. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10631–10638 (2020)
6. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1831–1840 (2017)
7. Ci, H., Wang, C., Ma, X., Wang, Y.: Optimizing network structure for 3d human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2262–2271 (2019)
8. Craig, J.J.: Introduction to robotics: mechanics and control, 3/E. Pearson Education India (2009)
9. Csiszar, A., Eilers, J., Verl, A.: On solving the inverse kinematics problem using neural networks. In: 2017 24th International Conference on Mechatronics and Machine Vision in Practice (M2VIP). pp. 1–6. IEEE (2017)
10. Girshick, R.; Radosavovic, I.G.G.D.P., Kaiming, H.: Detectron (2018), https://github.com/facebookresearch/detectron
11. Gong, K., Zhang, J., Feng, J.: Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8575–8584 (2021)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
13. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028 (2017)
14. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision, cambridge university press, new york, ny, usa, 2 edition (2003)
15. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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 (2014)
16. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: bmvc. vol. 2, p. 5. Citeseer (2010)
17. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7122–7131 (2018)

18. Kokic, M., Kragic, D., Bohg, J.: Learning to estimate pose and shape of hand-held objects from rgb images. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3980–3987. IEEE (2019)
19. Li, B., Li, X., Zhang, Z., Wu, F.: Spatio-temporal graph routing for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8561–8568 (2019)
20. Li, C., Lee, G.H.: Generating multiple hypotheses for 3d human pose estimation with mixture density network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9887–9895 (2019)
21. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3383–3393 (2021)
22. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3595–3603 (2019)
23. Li, S., Ke, L., Pratama, K., Tai, Y.W., Tang, C.K., Cheng, K.T.: Cascaded deep monocular 3d human pose estimation with evolutionary training data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6173–6183 (2020)
24. Liu, R., Shen, J., Wang, H., Chen, C., Cheung, S.c., Asari, V.: Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5064–5073 (2020)
25. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015)
26. Luo, C., Chu, X., Yuille, A.: Orinet: A fully convolutional network for 3d human pose estimation. arXiv preprint arXiv:1811.04989 (2018)
27. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recoverroboticsing accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 601–617 (2018)
28. 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)
29. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 international conference on 3D vision (3DV). pp. 506–516. IEEE (2017)
30. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. ACM Transactions on Graphics (TOG) **36**(4), 1–14 (2017)
31. Moon, G., Chang, J.Y., Lee, K.M.: Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10133–10142 (2019)
32. Mueller, F., Mehta, D., Sotnychenko, O., Sridhar, S., Casas, D., Theobalt, C.: Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1154–1163 (2017)

33. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7025–7034 (2017)
34. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7753–7762 (2019)
35. Peng, X., Tang, Z., Yang, F., Feris, R.S., Metaxas, D.: Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2226–2234 (2018)
36. Rogez, G., Schmid, C.: Mocap-guided data augmentation for 3d pose estimation in the wild. arXiv preprint arXiv:1607.02046 (2016)
37. Sharma, S., Varigonda, P.T., Bindal, P., Sharma, A., Jain, A.: Monocular 3d human pose estimation by generation and ordinal ranking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2325–2334 (2019)
38. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12026–12035 (2019)
39. Shi, M., Aberman, K., Aristidou, A., Komura, T., Lischinski, D., Cohen-Or, D., Chen, B.: Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. ACM Transactions on Graphics (TOG) **40**(1), 1–15 (2020)
40. Su, Z., Ye, M., Zhang, G., Dai, L., Sheng, J.: Cascade feature aggregation for human pose estimation. arXiv preprint arXiv:1902.07837 (2019)
41. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5693–5703 (2019)
42. Tang, W., Yu, P., Wu, Y.: Deeply learned compositional models for human pose estimation. In: Proceedings of the European conference on computer vision (ECCV). pp. 190–206 (2018)
43. Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V., Fua, P.: Structured prediction of 3d human pose with deep neural networks. arXiv preprint arXiv:1605.05180 (2016)
44. Tekin, B., Rozantsev, A., Lepetit, V., Fua, P.: Direct prediction of 3d body poses from motion compensated sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 991–1000 (2016)
45. Tripathi, S., Ranade, S., Tyagi, A., Agrawal, A.: Posenet3d: Learning temporally consistent 3d human pose via knowledge distillation. In: 2020 International Conference on 3D Vision (3DV). pp. 311–321. IEEE (2020)
46. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 109–117 (2017)
47. Wandt, B., Rosenhahn, B.: Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7782–7791 (2019)
48. Xu, J., Yu, Z., Ni, B., Yang, J., Yang, X., Zhang, W.: Deep kinematics analysis for monocular 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 899–908 (2020)

49. Zeng, A., Sun, X., Huang, F., Liu, M., Xu, Q., Lin, S.: Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In: European Conference on Computer Vision. pp. 507–523. Springer (2020)
50. Zhang, X., Xu, C., Tao, D.: Context aware graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14333–14342 (2020)
51. 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/CVF Conference on Computer Vision and Pattern Recognition. pp. 3425–3435 (2019)
52. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. arXiv preprint arXiv:2103.10455 (2021)