Supplementary Material
Weakly-Supervised Discovery of Geometry-Aware Representation for 3D Human Pose Estimation

Anonymous CVPR submission

Paper ID 2239

1. Introduction

This supplementary file presents: (1) the detail description of data augmentation strategies and network architecture used in our approach; (2) additional experimental analysis and quantitative results of our approach; (3) more qualitative evaluations of the effectiveness of our approach.

2. Implementation Details

In this section, we provide more details regarding the data augmentation strategies used in the training process, the network architecture of $generator(\cdot,\cdot)$, and the fusion mechanism used in the experiments for combining learnt geometry representation G with baselines.

Data Augmentation. We apply two types of data augmentation strategies to increase the diversity of the training samples. The first augmentation strategy is the *random in-plane rotations*. Given a ground-truth 3D pose $\mathbf{b}_{gt}$ and two randomly sampled viewpoints(i,j) with its relative rotation matrix$R_{i\to j}$, we rotate the 3D pose uniformly ($\pm 60°$) around z axe under one of (i,j) camera coordinates to obtain $\mathbf{b}_{aug}$. Then the new 2D pose training pair (S^i_{aug}, S^j_{aug}) with $R_{i\to j}$ are generated by projecting $\mathbf{b}_{aug}$ on perspective plane of cameras (i,j). The other data augmentation is based on *virtual cameras*. Similar to [1, 5], we synthesize the $\mathbf{b}_{aug}$ in virtual camera coordinate according to ground-truth 3D pose $\mathbf{b}_{gt}$. Then, we project $\mathbf{b}_{aug}$ on camera perspective plane to obtain 2D pose S_{aug}. Concretely, we assume all cameras roughly point towards a center position Q_{center}, which is the closest point to optical axes of all cameras provided by training set. Q_{center} is calculated by

$$Q_{center} = \arg\min_{Q} \sum_{a=1}^{A} d(Q, l_a), \tag{1}$$

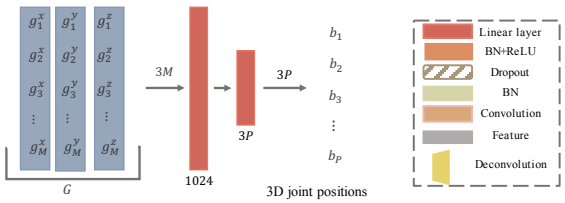
where l_a is the line indicates the optical axis, d is the distance between camera and Q_{center}. We sample d from a normal distribution with center and variance computed from the training set. Different from [1, 5] that generate new 2D coordinates-3D coordinates pair with sampling camera positions uniformly on a surface of a sphere, we synthesize 2D pose pair $S_{aug} = \{(S^i_{aug}, S^j_{aug})\}$ by randomly sampling two camera positions on the the torus with center Q_{center} and radius d.

	Layer	Module
Encoders	1	Conv-(N15,K4,S2,P1)
	2	LeaklyReLU, Conv-(N48,K4,S2,P1), BatchNorm
	3	LeaklyReLU, Conv-(N96,K4,S2,P1), BatchNorm
	4	LeaklyReLU, Conv-(N192,K4,S2,P1), BatchNorm
	5	LeaklyReLU, Conv-(N384,K4,S2,P1), BatchNorm
	6	LeaklyReLU, Conv-(N384,K4,S2,P1), BatchNorm
	7	LeaklyReLU, Conv-(N384,K4,S2,P1)
	8	Reshape(3,128,1)
	9	Multiply with Relative Rotation Matrix
Decoders	1	Reshape(384,1,1)
	2	ReLU, ConvT-(N384,K4,S2,P1), BatchNorm
	3	ReLU, ConvT-(N384,K4,S2,P1), BatchNorm
	4	ReLU, ConvT-(N192,K4,S2,P1), BatchNorm
	5	ReLU, ConvT-(N96,K4,S2,P1), BatchNorm
	6	ReLU, ConvT-(N48,K4,S2,P1), BatchNorm
	7	ReLU, ConvT-(N15,K4,S2,P1), Tanh

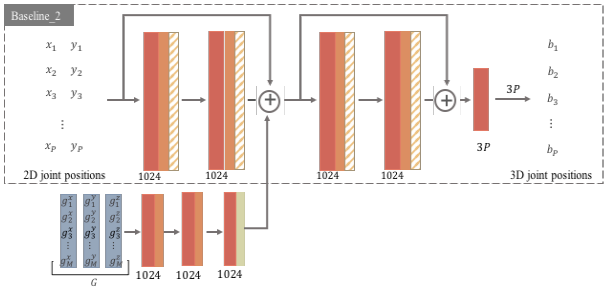
Table 1: Network architecture of the $generator(\cdot,\cdot)$. Conv represents the Convolutional layer, N denotes the number of channels, K denotes the kernel size, S denotes the stride size, and P denotes the padding size. ConvT corresponds to a layer performing transposed Convolution.

Network Architecture. The $generator(\cdot,\cdot)$ consists of an encoder and a decoder. The exact architecture is summarized in Table 1. We train the overall model to learn G for 90 epochs using Adam optimizer. The initial learning rate is 0.0001. The batch size is set to 64.

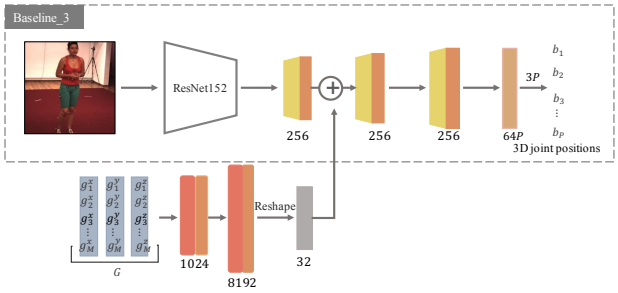
Fusion Mechanism. As mentioned in the main paper, we use three 3D pose estimators, *i.e.*, *Regression#1*, *Regression#2* and *Regression#3*, to evaluate the effectiveness of the learnt geometry representation G to 3D human pose estimation task. For *Regression#1*, the regression module is a two-layer fully-connected network. For the rest two regressors, in order to evaluate the robustness and flexibility of the proposed geometry representation in a



(a) Ours+Regression#1



(b) Ours+Regression#2



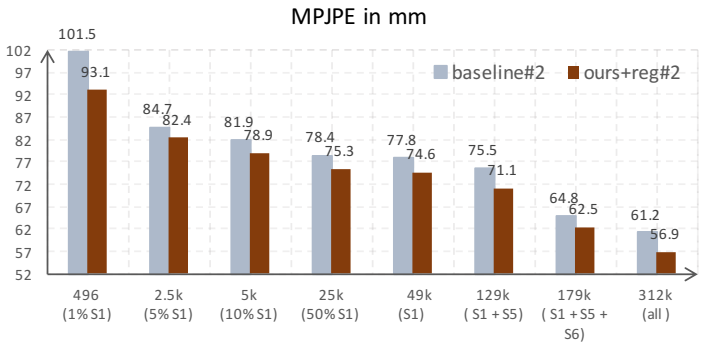
(c) Ours+Regression#3

Figure 1: The illustration of fusing the representation G with the original features of baselines (zoom in for more details).

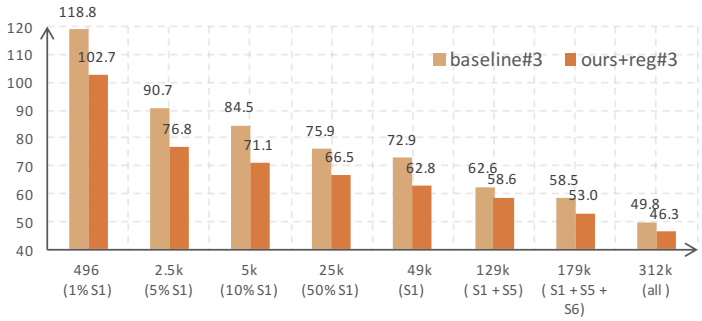
straightforward manner, we *only* forward the geometry representation G to fully connection layers to match the feature dimension of baselines, and then directly do element-wise sum with baselines [2, 4]. Baselines are trained with their public implementations and default settings. Figure 1 shows the detail positions of the fusion operation.

3. Additional Experimental Results

Results on Human3.6M under different amount of training samples. Besides the results analysed in the main paper, we also evaluate the effectiveness of the learnt representation G *as a robust 3D prior* to different 3D human pose estimation methods, on the condition of using the different amount of 3D annotated samples (under *Protocol#1*) to train the 3D pose estimators. As can be seen from Figure 2, our approach yields a consistent improvement over the baselines [2, 4] for all configurations. Specifically, given only about 500 annotated training samples (1%S1), the proposed G help reduce the errors for baseline#2 and



(a)



(b)

Figure 2: Evaluation on the Human3.6M using different number of training data under MPJPE metric.

baseline#3 by 8.3% ($101.5mm \rightarrow 93.1mm$) and 13.6% ($118.8mm \rightarrow 102.7mm$), respectively. Under the amount of 25000($50\%S1$) training samples, G boosts the baseline#2 to achieve better performance than single baseline#2 on larger amount of data ($49000(100\%S1)samples$), with particularly $2.5mm$ gains. For baseline#3, G help the model achieving $4.8mm$ gains with half training samples of S1. Table 2 reports the results of same configurations under PMPJPE metric, yielding similar observation. These phenomena confirm the flexibility and effectiveness of the proposed G to existing 3D human pose estimation methods, with improving the performance of these methods on less annotated training samples requisition.

Number of Training Data	Baseline#2	Ours+Reg#2	Baseline#3	Ours + Reg#3
496 (1%S1)	74.4	69.7	105.3	90.4
2.5k (5%S1)	62.3	60.5	82.6	69.3
5k (10%S1)	60	57.7	78.0	65.0
25k (50%S1)	57.5	55.5	73.6	62.4
49k (S1)	56.4	54.8	69.2	58.6
129k (S1+S5)	51.1	50.2	58.2	55.3
179k (S1+S5+S6)	49	47.6	55.6	50.2
312k (all)	47.7	44.1	44.1	41.6

Table 2: Evaluation on the Human3.6M using different number of training data under PMPJPE metric.

Ablation Study. To assess the effectiveness of the key com-

ponents of our approach to different 3D human pose estimators, except the ablation study results reported in the main paper, we also evaluate ablation studies on the configurations of $Regression\#1$ and $Regression\#2$. We clarify here that the baseline of "Ours+Reg#1" presents the performances of directly regressing 3D pose coordinates from 2D detections[1] with the same regressor (Regression#1). Table 3 presents the results for the two configurations. The level of relative improvements is varying on the configurations due to the effect of different network architectures and plain vanilla fusion mechanism on passing the valid information of our components to baselines. However, the tendencies are coherent. These ablation studies provide additional evidence that the key components of the proposed model are useful to the methods that injecting them and indeed robust to different configurations.

Components	Ours + Reg#1		Ours + Reg#2	
	MPJPE	$\triangle(\%)$	MPJPE	$\triangle(\%)$
BL	114.7	-	61.2	-
BL+SG	86.3	24.8$\downarrow$	57.6	6.3$\downarrow$
BL+SG+AUG	85.3	1.1$\downarrow$	57.5	0.18$\downarrow$
BL+DG+AUG	80.2	6.0$\downarrow$	56.9	1$\downarrow$

Table 3: Ablation studies on different components in our method with other baselines. The evaluation is performed on Human3.6M under *Protocol#1* with MPJPE metric. $\triangle$ is the relative error decrease.

4. Qualitative Evaluations

In this section, we show additional point cloud (*i.e.*, the manifold) interpolation to verify the robustness and $3D$ geometry semantics of the proposed representation G.

We show samples from the manifold, decoding them into 2D skeleton on the target domain and regressing them to 3D human pose with 'Ours+Regression#1'. Concretely, we randomly take two 2D poses (S_j^1 and S_j^2) under same camera viewpoint, encoding their corresponding source poses S_i^1 and S_i^2 to obtain latent samples G^1 and G^2. Then, the linear interpolation is applied between these two latent samples to obtain interpolated latent samples G^λ : $G^\lambda = \lambda G^1 + (1-\lambda)G^2$. We subsequently decode the G^λ into 2D skeleton and regress G^λ to 3D human pose, resulting in a triplet. For a straightforward visual perception, ***please refer to supplemental video*** to see the interpolation results. Note that, the 2D poses in the first and last frames are S_j^1 and S_j^2. The rest results on the video are synthesized results. As can be seen from the video, the results are consistent amongst 2D skeleton and 3D poses under the changing of latent samples. This shows the proposed representation G has extracted semantic 3D geometry representation of the human pose. Moreover, the smooth interpolation results show that a reasonable coverage of the manifold has been successfully learned by our model, yielding a robust geometry representation to diverse poses.

[1]The 2D detections are obtained from a pre-trained 2D human pose estimator [3].

References

[1] H. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *AAAI*, 2018. 1

[2] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 2

[3] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 3

[4] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei. Integral human pose regression. In *ECCV*, 2018. 2

[5] M. Wang, X. Chen, W. Liu, C. Qian, L. Lin, and L. Ma. Drpose3d: Depth ranking in 3d human pose estimation. *arXiv preprint arXiv:1805.08973*, 2018. 1