Residual Deep Monocular 3D Human Pose Estimation using CVAE synthetic data

Yutong Jiang¹*, Xiaofei Liu², Dongyuan Wu³, Pengbiao Zhao⁴

¹ School of Information and Software Engineering, UESTC, Chengdu, Sichuan, 610000, China,
² Daqing Drilling & Exploration Engineering Corp. NO.1 Drilling company Market Management Center Iraq Rumaila project, Daqing, Heilongjiang, 163000, China
³ School of Information and Software Engineering, UESTC, Chengdu, Sichuan, 610000, China
⁴ School of Information and Software Engineering, UESTC, Chengdu, Sichuan, 610000, China

*e-mail: czyang@std.uestc.edu.cn

Abstract. Estimating 3D human pose from a monocular RGB image is a challenging task in computer vision because annotating a large number of 3D pose ground-truth data is costly. To address the problem of lack of 3D data, many methods have been proposed. In this paper, we propose to address this issue by synthesizing data. Firstly we exploit Conditional Variational Autodecoder to generate 3D human skeletons data. In CVAE network, we acquire generous 3D pose samples through the predicted 2D pose and the existing 3D ground truth. Secondly, base the generated 3D samples, we obtain the corresponding 2D pose by projection, thus augmenting the data of 2D-3D network. Finally, we train these synthetic data on 3D pose residual estimation network. Extensive experiments show that our approach achieves state-of-the-art accuracy on standard benchmark datasets.

1. Introduction

Human pose estimation is a hot research in computer vision, and has widely application, such as movie animation, virtual reality, human-computer interaction, intelligent monitoring, athlete assisted training. With the development of neural network and deep learning, the end-to-end prediction network has been applied on human pose estimation and made remarkable progress[1][2][3][4][5]. Particularly, 3D human pose estimation from single image has attracted extensive attention and achieved unexpected success[6][7][8][9]. Although these methods have achieved noteworthy accuracy, 3D human pose estimation is still a huge challenge since 3D data sets are limited and collecting 3D annotations is costly and time-consuming.

To deal with this problem, many recent works adopt a two-stage network framework, which first predicts 2D joint points, and then lifts to 3D[10][11][12]. 3D human pose estimation task can obtain more extra features after 2D pose detector since 2D annotation is easier to obtain and more diverse. However, there is still a big difference between the results of the test set and the training set due to the data bias and the limitation of 3D data. To solve this problem, we use CVAE to synthesize 3D joint
samples. The corresponding 2D joint points are obtained by using the mapping relation between 3D and 2D bone. We performed ablation experiments to demonstrate the effectiveness of our method.

2. Related Works

**Two-step pose estimation.** 3D coordinates are obtained by direct regression from 2D images\cite{13}\cite{14}\cite{15}. Due to the high degree of non-linearity of 3D space and large output space, the features that need to be learned by a single model are too complex. Usually a satisfactory result is not achieved. Thanks to the maturity of 2D pose estimation, lifting 2D to 3D achieved unexpected accuracy\cite{16}\cite{17}\cite{18}\cite{19}. Two-step pose estimation has two branches. One is joint 2D and 3D training, the other is directly using the pre-trained 2D pose network to input the 2D coordinates into the 3D pose estimation network. Our work belongs to the second branch. Since HRNet can effectively extract appearance information\cite{20}\cite{21}. In this paper, we use HRNet as 2D pose detector to get 2D joint points as input for next stage. HRNet maintains high-resolution representation by connecting high-resolution convolutions to low resolution convolutions in parallel, and enhances high-resolution representation by repeatedly performing multi-scale fusion across parallel convolutions.

**Self-supervised/weakly methods.** Network training needs a large number of data, otherwise it is easy to overfit, but 3D annotation is often difficult to obtain. To address this problem, Self-supervised and weakly methods were proposed\cite{22}. Chen x, et al\cite{23}. proposed a novel weakly supervised encoder decoder framework to learn geometric perceptual 3D representation of human posture. In order to improve the robustness of representation geometry, the loss of representation consistency is introduced to constrain the learning process of encoder decoder. Our method also uses decoder and encoder. But the difference is that we use it to generate 3D bone samples. Recently, multi-perspective self-supervised approach has been proposed\cite{24}. 3D pose is obtained by 2D pose estimation and epipolar geometry to reduce the dependence on 3D ground-truth. Considerable results are obtained by two views.

**Monocular 3D Human Pose Estimation.** In a real world scenario, there are many cases where only one view of the data set is available. In the early stage, the paper\cite{25} is the first one to apply deep learning to 3D pose estimation based on a single image. It designs a deep convolution network to directly regress 3D coordinates from RGB images, and combines the regressed key point coordinates with the detected key point bounding box by means of multi task learning. In the supervision, instead of using 3D coordinates directly, they use the backbone length and the key point bounding box to supervise. Pavlakos G, et al\cite{14}. used the method of 2d pose in 3D to regress a 3D Heatmap. Considering a large range of z-axis depth, they adopted a coarse to fine structure to regress step by step. It is similar to the stacked hourglass structure in 2D pose estimation. Unlike these methods, we focus on data enhancement.

![Figure 1. Overall architecture of predicting 3D pose](image-url)
3. Models and Methods

As illustrated in Figure 1, our overall architecture consists of three parts: CVAE data generator, 3D-2D projection and 3D pose estimation network. CVAE data generator is a encoder-decoder network that generate 3D joint samples. 3D-2D projection maps the generated 3D joint points to corresponding 2D joint points. 3D pose estimation network predicts 3D pose through 2D joint points.

3.1. CVAE data generator

CVAE is one of the most advanced methods to generate models, which has achieved remarkable results[26]. CVAE is widely used in computer vision. It includes an encoder and a decoder. The encoder attempts to learn the hidden representation or probabilistic encoder representation of the data. The decoder attempts to learn to hide the representation input space. We use CVAE to generate N 3D joint samples. The 3D ground truth(\(p_{3D}\)) and the 2D joint points(\(\tilde{p}_{2D}\)) predicted by HRNet are used as the input of the network. First, we compute a prior distribution \(q(z)\). The encoder obtains the hidden representation \(p_{3D} = p_{3D}(\tilde{z})\) by learning the features of \(p_{3D}\) and \(\tilde{p}_{2D}\). The decoder obtains \(\hat{p}_{3D} = \hat{p}_{3D}(n = 1, 2, ..., N)\) by decoding the hidden representation. We optimize CVAE network by back propagation. The loss function consists of two parts: one is reconstruction loss, the other is KL divergence between prior distribution and conditional data distribution. We set the corresponding weights for the two losses(\(\alpha, \beta\)). We compute loss as follows:

\[
Loss1 = E_{z \sim q(z)} || \hat{p}_{3D} - P_{3D} ||^2_2 \\
Loss2 = KL(q(\tilde{z}) | p_{3D}, \tilde{p}_{2D}) || p(z | \tilde{p}_{2D}) \\
L_{CVAE} = \alpha Loss1 + \beta Loss2
\]

3.2. 3D-2D projection

Perspective projection is a method of drawing or rendering on two-dimensional image or canvas in order to obtain the visual effect close to real three-dimensional objects. Perspective projection maps the generated 3D joint points to corresponding 2D joint points. The coordinates of each 3D joint point and each 2D joint point are respectively expressed as \((x_i, y_i, z_i)\), \((a_i, b_i)\). Mapping 2D coordinates by focal length \(f = \{f_x, f_y\}\) and point\((c = \{c_x, c_y\})\). The specific format is as follows:

\[
\begin{align*}
a_i &= \frac{x_i}{z_i} f_x + c_x \\
b_i &= \frac{y_i}{z_i} f_y + c_y
\end{align*}
\]

3.3. 3D pose estimation network

After obtaining a large number of 2D and 3D joint points, the prediction network is trained to obtain 3D pose. The network consists of several block. These block is composed of full connection and residual network as illustrated in Figure 2. The input of the first layer is 2D joint points, and the input of the subsequent layer is the contact of 2D joint predicted by HRNet and predicted 3D joint points(\(\tilde{p}_{3D}\)).

4. Experiments

4.1. Datasets & Evaluation Metrics

We conduct experiments on Human3.6M that is the largest dataset for 3D human estimation. We use subject 1 to 5 as training set, 9 and 11 as verification set. We evaluate our model by mean per joint
Figure 2. 3D pose estimation network

position error (MPJPE) denoted as P1 and MPJPE after alignment involving rotation and translation (PA-MPJPE) as P2.

4.2. Results

We performed extensive experiments to demonstrate the effectiveness of our method. The results showed that the average error rate and the error rate of part of actions under P1 and P2. Table 1 and Table 2 show the comparison of our experimental results with other methods.

Table 1. Results of our model on Human3.6M under Protocol 1

|     | Direction | Discussion | Eating | Pose | Sitting | Smoking | Wait | Walk | Avg  |
|-----|-----------|------------|--------|------|---------|---------|------|------|------|
| Pavlakos et al. (2017) [14] | 67.4 | 71.9 | 66.7 | 65.0 | 83.7 | 71.7 | 65.8 | 59.1 | 114.1 |
| Fang et al. (2018) [17]   | 50.1 | 54.3 | 57.0 | 53.4 | 72.8 | 60.3 | 57.7 | 47.5 | 60.4 |
| Zhao et al. (2019) [27]   | 54.8 | 60.7 | 58.2 | 53.8 | 75.2 | 64.1 | 66.0 | 63.2 | 64.9 |
| Liu et al. (2020) [28]    | 50.7 | 60.0 | 51.1 | 48.8 | 72.8 | 58.6 | 61.0 | 45.9 | 61.1 |
| Xia et al. (2020) [29]    | 49.6 | 54.3 | 53.5 | 51.2 | 64.7 | 56.8 | 53.4 | 44.3 | 56.7 |
| Ours                       | 45.8 | 47.2 | 49.6 | 47.4 | 57.3 | 49.6 | 48.4 | 41.5 | 51.1 |

Table 2. Results of our model on Human3.6M under Protocol 2

|     | Direction | Discussion | Eating | Pose | Sitting | Smoking | Wait | Walk | Avg  |
|-----|-----------|------------|--------|------|---------|---------|------|------|------|
| Fang et al. (2018) [17] | 38.2 | 41.7 | 43.7 | 40.2 | 54.5 | 47.2 | 44.3 | 37.7 | 45.7 |
| Drover et al. (2018) [30] | 60.2 | 60.7 | 59.2 | 59.4 | 69.1 | 64.8 | 60.8 | 63.9 | 64.6 |
| Xia et al. (2020) [29]   | 39.0 | 43.4 | 42.4 | 42.3 | 54.5 | 46.4 | 41.4 | 34.4 | 45.1 |
| Ours                       | 34.7 | 36.3 | 36.9 | 35.5 | 42.2 | 38.2 | 36.9 | 30.9 | 38.9 |

5. Conclusions

In this paper, we propose to use CVAE to generate 3D skeleton data and use perspective projection to get the corresponding 2D joint points, so as to increase the data of 3D prediction network. We effectively reduce the limitation of dataset. Extensive experiments show that the proposed method achieves close to state-of-the-art results on benchmark datasets.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No.62072074, No.62076054, No.62027827, No.61902054), the Frontier Science and Technology Innovation Projects of National Key R&D Program (No.2019QY1405), the Sichuan Science and Technology Innovation
Platform and Talent Plan (No.2020JDJQ0020), and the Sichuan Science and Technology Support Plan (No. 2020YFSY0010).

References
[1] Cheng, Y., Yang, B., Wang, B., Wending, Y., & Tan, R. (2019). Occlusion-Aware Networks for 3D Human Pose Estimation in Video. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE.

[2] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun(2018). Cascaded pyramid network for multi-person pose estimation. In CVPR, pages 7103–7112.

[3] Kwabena, O. A., Qin, Z., Zhuang, T., & Qin, Z. (2019). Mscryptonet: multi-scheme privacy-preserving deep learning in cloud computing. IEEE Access.

[4] Zhen Qin, Gu Huang, Hu Xiong, Zhiguang Qin, and Kim-Kwang Raymond Choo.(2019) A Fuzzy Authentication System Based on Neural Network Learning and Extreme Value Statistics. IEEE Transactions on Fuzzy Systems.

[5] Alejandro Newell, Kariyu Yang, and Jia Deng(2016). Stacked hour glass networks for human pose estimation. In ECCV.

[6] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis.(2017). Coarse-to-fine volumetric prediction for single-image 3d human pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7025–7034.

[7] Xiao Sun, Bin Xiao, Fangying Wei, Shuang Liang, and Yichen Wei. Integral human pose regression(2018). In Proceedings of the European Conference on Computer Vision (ECCV), pages 529–545.

[8] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning(2018). In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5137–5146.

[9] Mude Lin, Liang Lin, Xiaodan Liang, Keze Wang, and Hui Cheng. Recurrent 3d pose sequence machines(2017). In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 810–819.

[10] Francesc Moreno-Noguer.(2017). 3d human pose estimation from a single image via distance matrix regression. In CVPR, pages 2823–2832.

[11] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Saifeer Afaque, Abhishek Sharma, and Arjun Jain.(2018). Learning 3d human pose from structure and motion. In ECCV.

[12] Chen, C.-H., RamananD. (2017). 3D human pose estimation = 2D pose estimation + matching. In CVPR. 7035–7043.

[13] Li, S., & Chan, A. B. (2014). 3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network. Asian Conference on Computer Vision. Springer, Cham.

[14] Pavlakos, G., Zhou, X., Derpanis, K. G., & Daniilidis, K. (2017). Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. IEEE Conference on Computer Vision & Pattern Recognition. IEEE.

[15] Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., & Cai, J., et al. (2020). 3D Hand Shape and Pose Estimation From a Single RGB Image. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.

[16] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis.(2016). Sparseness meets deepness: 3d human pose estimation from monocular video. In CVPR, pages 4966–4975.

[17] Hao-Shu Fang, Yuanlu Xu, Wenguang Wang, Xiaobai Liu, and Song-Chun Zhu.(2018) Learning pose grammar to encode human body configuration for 3d pose estimation. In AAAI.

[18] Zeng, A., Sun, X., Huang, F., Liu, M., Xu, Q., & Lin, S. (2020). SRNet: Improving Generalization in 3D Human Pose Estimation with a Split-and-Recombine Approach.
[19] Chen, Z., Huang, Y., Yu, H., Xue, B., Han, K., & Guo, Y., et al. (2020). Towards Part-Aware Monocular 3D Human Pose Estimation: An Architecture Search Approach. European Conference on Computer Vision. Springer, Cham.

[20] Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. arXiv e-prints.

[21] Zhen Qin, Wei He, Fuhu Deng, Meng Li, and Yao Liu. (2019). SRPRID: Pedestrian Re-identification based on Super-resolution Images. IEEE Access, Volume 7, Issue 1: 152891-152899.

[22] Dylan Drover, Rohith MV, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh. (2018). Can 3d pose be learned from 2d projections alone? European Conference on Computer Vision Workshops.

[23] Chen, X., Lin, K., Liu, W., Qian, C., and Lin, L. (2019). Weakly-Supervised Discovery of Geometry-Aware Representation for 3D Human Pose Estimation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 10887-10896.

[24] Kocabas, M., Karagoz, S., & Akbas, E. (2019). Self-Supervised Learning of 3D Human Pose Using Multi-View Geometry. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.

[25] Li, S., & Chan, A. B. (2014). 3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network. Asian Conference on Computer Vision. Springer, Cham.

[26] Du, L., Ding, X., Liu, T., & Li, Z. (2019). Modeling Event Background for If-Then Commonsense Reasoning Using Context-aware Variational Autoencoder.

[27] Zhao, L., Peng, X., Tian, Y., Kapadia, M., and Metaxas, D. N. (2019). Semantic Graph Convolutional Networks for 3D Human Pose Regression. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 3420-3430.

[28] Liu, J., Ding, H., Shahroudy, A., Duan, L. Y., & Kot, A. C. (2020). Feature boosting network for 3d pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), 494-501.

[29] Xia, H. and Xiao, M. (2020). 3D Human Pose Estimation With Generative Adversarial Networks. in IEEE Access, vol. 8, pp. 206198-206206.

[30] Drover, D., Chen, C.-H., Agrawal, A., Tyagi, A., and Huynh, C. P. (2018). Can 3D pose be learned from 2D projections alone? in Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 78–94.