Unsupervised 3D Human Pose Estimation in Multi-view-multi-pose Video

Cheng Sun
Kyushu University
Email: sun.cheng.736@s.kyushu-u.ac.jp

Diego Thomas
Kyushu University
Email: thomas@ait.kyushu-u.ac.jp

Hiroshi Kawasaki
Kyushu University
Email: kawasaki@ait.kyushu-u.ac.jp

***Abstract*—3D human pose estimation from a single 2D video is an extremely difficult task because computing 3D geometry from 2D images is an ill-posed problem. Recent popular solutions adopt fully-supervised learning strategy, which requires to train a deep network on a large-scale ground truth dataset of 3D poses and 2D images. However, such a large-scale dataset with natural images does not exist, which limits the usability of existing methods. While building a complete 3D dataset is tedious and expensive, abundant 2D in-the-wild data is already publicly available. As a consequence, there is a growing interest in the computer vision community to design efficient techniques that use the unsupervised learning strategy, which does not require any ground truth 3D data. Such methods can be trained with only natural 2D images of humans. In this paper we propose an unsupervised method for estimating 3D human pose in videos. The standard approach for unsupervised learning is to use the Generative Adversarial Network (GAN) framework. To improve the performance of 3D human pose estimation in videos, we propose a new GAN network that enforces body consistency over frames in a video. We evaluate the efficiency of our proposed method on a public 3D human body dataset.**

I. Introduction

3D human pose estimation is a fundamental problem in computer vision aiming to extract 3D poses of people from 2D images or videos. It has many applications in areas such as motion capture, surveillance, robot technology, computer generated imagery, medical science and military. The objective is to estimate the relative 3D orientation and length of bones (up to an unknown scale) of a template human skeleton from in-the-wild 2D observations. This is known to be an ill-posed problem due to the multiple possible combinations of pose, shape, color and illumination that can produce the exact same 2D image. Nevertheless, with the help of deep learning and constraints carefully designed for the human body, several solutions have recently been proposed with impressive results [1]–[3].

Recently, using fully supervised methods have been proven to be a successful strategy on publicly available datasets. Effective approaches have been proposed that focus on multi-view pose estimation or pose estimation from videos. However, most of existing methods require to train the 3D human pose estimation networks with labeled 3D ground truth data, which are few. As a consequence, state-of-the-art techniques lack in generalizability. To generate ground truth 3D pose dataset, expensive equipment such as motion capture systems must be used. Such systems need to be carefully calibrated and require an elaborate setup with multiple sensors and bodysuits, which is impractical to use outside.

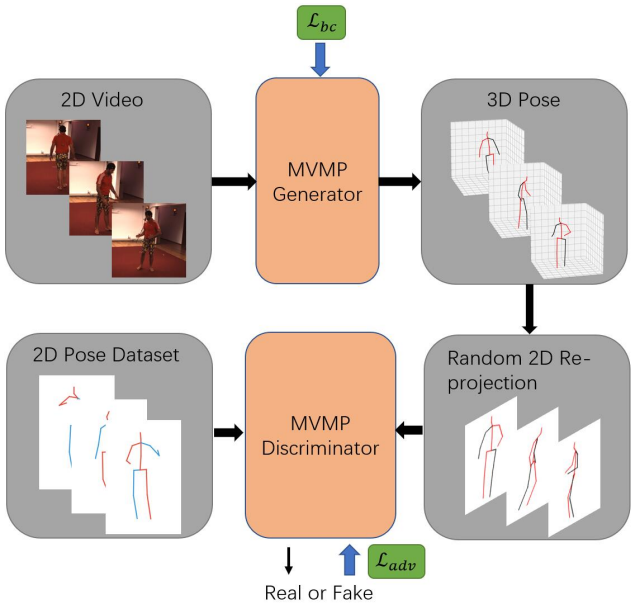

Fig. 1. Overview of our model. Our proposed method takes as input a multi-view-multi-pose (MVMP) 2D video and outputs the corresponding set of 3D poses using an unsupervised learning strategy.

On the other hand, there is abundant unlabeled in-the-wild 2D data such as videos on YouTube or photos on Instagram. How to take full advantage of these abundant unlabeled 2D data for 3D human pose estimation is a challenging problem which is attracting growing interest in the Computer Vision Community. Weakly-supervised methods use partially labeled data or data with low reliability as input. Unsupervised methods do not need any labeled data when training the network.

While most 3D human pose estimation methods from videos are based on fully supervised training, there are few researches about using unsupervised methods to estimate 3D poses from videos. Inspired by the work of Kudo *et al*. [4], we extend the unsupervised single-frame 3D pose estimation framework for the case of multi-frame 3D pose estimation. We notice that although multi-view 3D pose estimation has been popular and successful, all the existing works assume that the person is in the same pose from a single view or different views. In contrast we argue that the pose of the person is changing in different

views in the case of a video. This is a new challenging problem that we call multi-view-multi-pose (MVMP) estimation. Our proposed method uses the GAN framework as a backbone method for unsupervised 3D human pose estimation from 2D input videos. To take full advantage of the input videos, we introduce body consistency in the structure by considering the consistency of human skeletons over several frames which is shown in Fig.1. Compared to the original work of [4], our proposed method can reduce the error by 10%. The main contributions of our work are two-fold:

- We extend our model to process multi-frame 3D human pose estimation with input of videos.
- To improve the accuracy we introduce constraints of body consistency over frames into our model and reduce the error effectively.

II. Related Work

Existing 3D pose estimation methods can be divided into end-to-end methods and 2-step methods. It is proven that 2-step methods which separate the task of 3D pose estimation into 2D pose estimation and lifting 2D pose to 3D subsequently outperform end-to-end approaches [1]. We focus our discussion about related works on methods that use the 2-step approach. We further separate existing works into three classes: (A) fully supervised methods, (B) Weakly supervised methods and (C) unsupervised methods.

Recently general adversarial network(GAN) is widely used to implement weakly supervised methods and unsupervised methods [2], [5]. Given a training dataset, GAN learns to generate new data with same features as the training dataset [6]. A typical GAN always consists of a generator and a discriminator. The generator tries to produce images similar to the input images and confuse the discriminator. The discriminator tries to distinguish the output of generator from real images in the dataset.

A. Fully supervised training

Fully supervised methods have been proposed that use paired ground truth 2D-3D data for training. The 2D data consists of ground truth 2D locations of joint landmarks, while the 3D data consists of the corresponding ground truth 3D coordinates of the joints. For example, Martinez *et al*. [1] propose to learn a regression network to estimate 3D joints from 2D joints. Moreno-Noguer *et al*. [7] propose to learn a regression network from 2D distance matrix to 3D distance matrix by using 2D-3D correspondences. Exemplar based methods [8]–[10] are proposed which use 3D skeletons for nearest-neighbor look-up. Tekin *et al*. [11] combined 2D and 3D image cues relying on 2D-3D correspondences. Wang *et al*. [12] use 3D data to train an intermediate ranking network and extract the depth ordering of pairwise human joints from a single image. Sun *et al*. [13] use a 3D regression network based on bone segments instead of using joint coordinates directly.

B. Weakly supervised training

Weakly supervised approaches use unpaired 3D data to learn priors on shapes or poses. Part of data used in weakly supervised approaches can be unlabeled or with low-quality labels. For example, Zhou *et al*. [14] use a 3D pose dictionary and capture how poses appear from different camera views. Brau *et al*. [15] treat 3D pose as hidden variable and learn pose priors with an independently trained network. Tome *et al*. [16] leverage 2D annotations to train networks for estimating 3D poses. Tung *et al*. [17] propose to use an adversarial network with 2D projection consistency to learn from unpaired 2D-3D data. Ronchi *et al*. [18], use 2D data with ordinal annotations for weakly supervised 3D estimation. Dabral *et al*. [19] propose to implement additional constrains on estimated bone length and joint angles. Rhodin *et al*. [20] utilize multi-view consistency constraints with a few 3D ground truth data to avoid poses collapsing to a single location. Yang *et al*. [2] propose a multi-source adversarial structure which takes advantage of in-the-wild data to improve the performance of 3D pose estimation.

C. Unsupervised training

Rhodin *et al*. [21] implement an unsupervised approach to learn a geometry-aware body representation. They propose an encoder-decoder to map one view of human pose to another view with multi-view consistency loss. Kudo *et al*. [4] propose an unsupervised adversarial structure with re-projection constraint to recover 3D pose estimation with unlabeled 2D data. They have better results compared with baseline approach [22]. Chen *et al*. [5] propose to add geometric self-supervision to improve the performance of the unsupervised adversarial network. We are inspired by this strategy, but we focus our work on 3D pose estimation from videos while the methods mentioned above focus on estimation from single frames.

D. Pose estimation in video

Most of previous works concentrate on estimation from single frames. Recently there have been efforts in implementing estimation from videos taking advantage of temporal information to produce accurate estimation which is robust to noise. B. Tekin *et al*. [23] propose to infer 3D poses from the histograms of oriented gradients (HoG) features of spatio-temporal volumes. Katircioglu *et al*. [24] use LSTMs to refine 3D poses predicted from single images because of the well-known efficiency of LSTM for the task of temporal input. Hossain *et al*. [25] propose a sequence-to-sequence LSTM structure which encodes a sequence of 2D poses from a video and then decodes it into 3D poses. Lee *et al*. [26] propose to use RNN to learn 3D estimation from priors on body part connectivity. Recently Pavllo *et al*. [3] propose a temporal dilated convolutional model which can produce robust 3D pose estimation from videos whose length can be up to 243 frames.

III. Our Proposed Method

We propose a multi-view-multi-pose (MVMP) 3D pose estimation method using unsupervised adversarial structure.

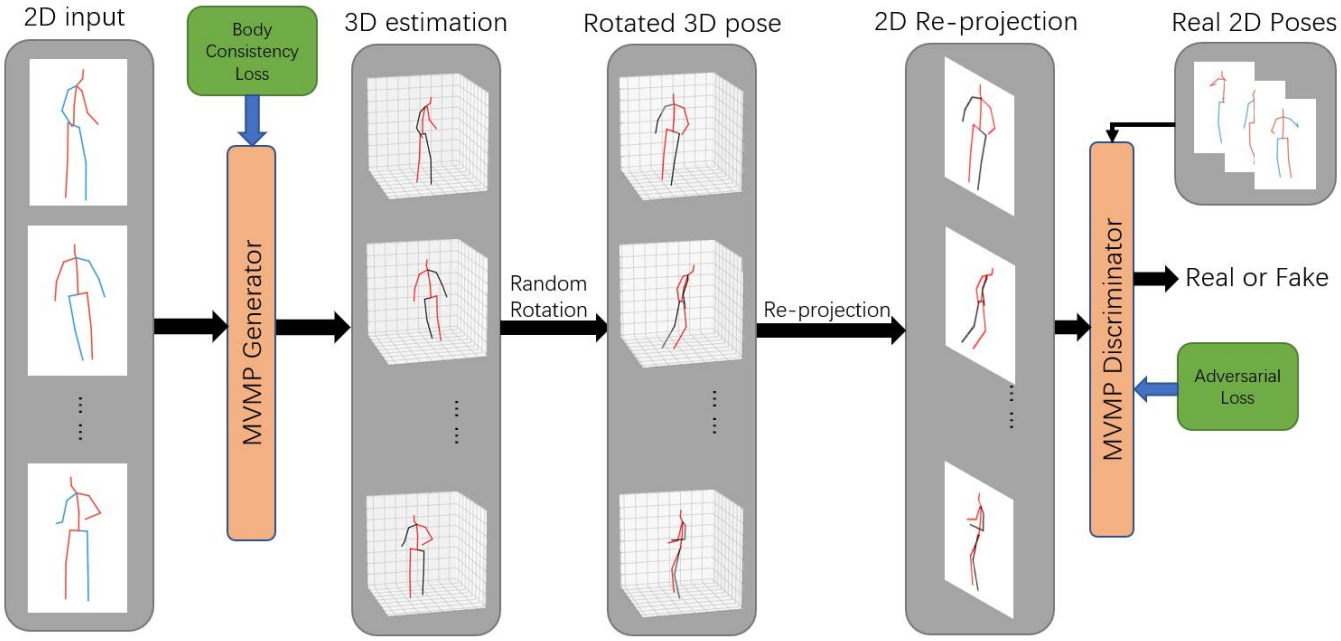

Fig. 2. Framework of our model. After the 2D keypoints are fed into generator, 3D pose estimation will be produced. Body consistency is calculated by the bone length differences of 3D skeletons to optimize the generator network. The 3D pose will be fed into discriminator. After rotated randomly and re-projected to 2D plane, adversarial loss will be calculated by difference of re-projection and real 2D images to further optimize the network.

Body consistency is implemented in the model to improve the performance with multi-frame input.

A. Overview

As shown in Fig.2, our proposed model consists of a generator and a discriminator similar to a typical GAN. A sequence of images sampled from the input video is fed into the generator. The generator in the network works as a 3D pose estimator. The outputs of the generator are joint depths of corresponding input 2D joints. Therefore 3D joint locations can be computed from 2D joint locations and estimated joint depths. The 3D pose will be rotated by random degree while the spin axis is perpendicular to the ground, since the camera views are usually parallel to the ground plane. The rotated 3D pose is then re-projected orthogonally into the 2D image plane. The re-projection is fed to the discriminator and the discriminator learns to distinguish the 2D-projection from real 2D poses which are in the dataset. During the training process, the generator learns to produce plausible 3D poses and corresponding 2D re-projections to confuse the discriminator, and the discriminator tries to discriminate between the re-projected 2D poses and the real ones. If the 3D poses given by the generator is plausible enough, it will be hard for the discriminator to distinguish the re-projection from a real 2D pose. Therefore a well-trained generator can be considered as a good 3D pose estimator after the training process. We introduce a network structure similar to [1] [5] in which the generator consists of four residual blocks and the discriminator consists of three residual blocks.

B. Adversarial Loss

The balancing problem between generator and discriminator during training process is one of the well-known problems of GAN. For example the discriminator sometimes performs better than the generator, which makes the generator perform much worth than expected. The loss function used in backbone method [4] is similar to the loss function proposed by Goodfellow *et al*. [6] :

$$\mathcal{L}_{adv} = E_{x\sim P_{data}}[\log D(x)] + E_{x\sim P_G}[\log(1 - D(x))] \quad (1)$$

where P_{data} refers to the probability distribution of real 2D pose, P_G refers to the probability distribution of generated 2D pose and D refers to the discriminator. When the network is being trained, the generator aims to minimize $\mathcal{L}_{adv}$ while the discriminator aims to maximize $\mathcal{L}_{adv}$.

One of the current solutions to solve the balancing problem of original GAN is Wasserstein GAN using Wasserstein distance [27], which is a measure of the distance between two probability distributions known as Earth Mover's distance. Wasserstein GAN, also known as WGAN, uses Wassterstein loss to replace the loss function in the original GAN. The loss function is shown as follows:

$$\mathcal{L}_{adv} = E_{x\sim P_{data}}[f_w(x)] - E_{x\sim P_G}[f_w(x)] \quad (2)$$

where f_w refers to a discriminator network with parameters w. w is limited in a certain range when the discriminator is being trained. Wasserstein GAN allows significant improvements compared with the GAN, notably on solving the balancing problem of GAN.

C. Sampling Strategies

The input of our proposed GAN is an image sequence with a length of L frames sampled from the input video. We compared two sampling strategies. The first strategy samples adjacent frames in the video and the second strategy samples frames between an interval of L/k to take more advantage of the temporal information hidden in the video.

D. Body Consistency

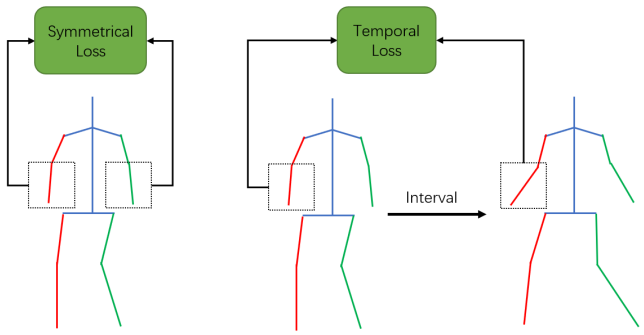

Fig. 3. Illustration of body consistency. The body consistency includes two parts, one is the difference between symmetric body part within a single skeleton, the other is the difference between the corresponding body parts over several frames.

We propose to use loose body consistency constraints to improve the accuracy of 3D pose estimation from 2D videos as shown in Fig.3. These constraints enforce the bone length symmetry between different body parts of a single person and bone length of the same person in different frames.

In a single frame we reason that the corresponding parts of a single person, such as left arm and right arm, should have same bone lengths.

$$\mathcal{L}_{sym} = \sum_{i=1}^{k} \sum_{j \in S_{sym}} \left\| B_{ij} - B'_{ij} \right\|_2^2 \tag{3}$$

where k is the number of frames in the input sample, and i is the frame number. S_{sym} includes symmetric body parts such as left arm and left leg. B_{ij} is the bone length of a body part such as left arm and B'_{ij} is the bone length of the symmetrical body part of B_{ij} such as right arm.

In multiple frames we argue that the length of the bones of the same body part of a person should not change.

$$\mathcal{L}_{tem} = \sum_{i=1}^{k-1} \sum_{j \in S_{tem}} \left\| B_{ij} - B_{(i+1)j} \right\|_2^2 \tag{4}$$

where S_{tem} includes all the body parts. B_{ij} is the bone length of j-th body part in i-th frame.

IV. Experiments

We show quantitative results and qualitative results on the widely used Human3.6M dataset for evaluation. During the training process we assume that 3D ground truth is unavailable. Only in testing process we use 3D ground truth for evaluation purpose.

A. Dataset and Evaluation

Human3.6M is one of the largest 3D human pose datasets, consisting of 3.6 million 3D human poses. The dataset contains video and motion capture (MoCap) data from 5 female and 6 male subjects. Data is captured from 4 different viewpoints, while subjects perform typical activities such as talking on the phone, walking, eating, etc. In our experiment we use subjects 1,5,6,7,8 for training and subjects 9,11 for testing. To evaluate our system, we use the Mean Per Joint Position Error(MPJPE) which is widely used in 3D human pose estimation.

B. Training details

We used a batch size of 512 which balanced the speed and performance. We trained the network for 50 epochs. Following Kudo *et al.* [4], the weight of each layer was initialized by Gaussian distribution with standard deviation of 0.14. The generator and discriminator were updated in every epoch. To optimize the balance between generator and discriminator we stopped updating the generator when its accuracy became larger than 0.9 and stopped updating discriminator if the accuracy became smaller than 0.3. The model was trained with Intel Core i7-8700K CPU, Nvidia 1080Ti GPU and 16GB RAM.

C. Results and Discussions

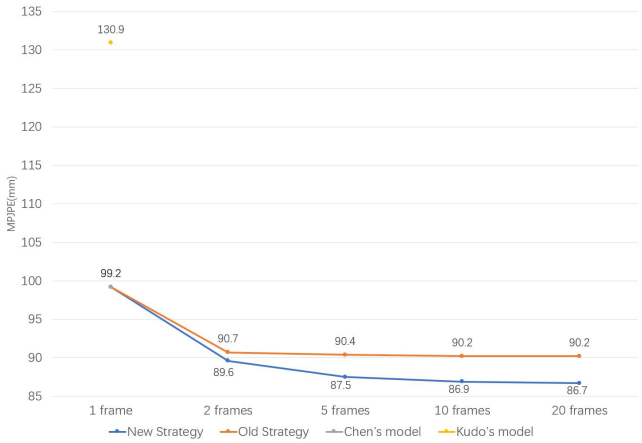

Fig. 4. Quantitative results. The comparison between our model with different sampling frame amounts and sampling strategies and single-image models.

As shown in Table 1 we compared our results with state-of-the-art results including methods processing single images and videos. Our approach with loose body consistency constraints outperformed most unsupervised methods. Note that the implementation of the network proposed by Chen *et al.* [5] is not provided by the authors. As a consequence we did our best to faithfully implement the method in [5] based on the descriptions in the paper. We report the results obtained with our own implementation of [5]. Even though the results obtained with our proposed method are not as good as those obtained by state-of-the-art supervised methods, our proposed method does not rely on any 3D ground truth dataset.

TABLE I
QUANTITATIVE RESULTS OBATAINED ON THE HUMAN3.6M DATASET

Method	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
supervised																
Martinez *et al*. [1]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Iskakov *et al*. [28]	19.9	20.0	18.9	18.5	20.5	19.4	18.4	22.1	22.5	28.7	21.2	20.8	19.7	22.1	20.2	20.8
self-supervised																
Kocabas *et al*. [29]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	60.6
semi-supervised																
Pavllo *et al*. [3]	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
weakly-supervised																
Wandt *et al*. [30]	77.5	85.2	82.7	93.8	93.9	101.0	82.9	102.6	100.5	125.8	88.0	84.8	72.6	78.8	79.0	89.9
Yang *et al*. [2]	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	43.6	60.1	47.7	58.6
unsupervised																
Kudo *et al*. [4]	125.0	137.9	107.2	130.8	115.1	127.3	147.7	128.7	134.7	139.8	114.5	147.1	130.8	125.6	151.1	130.9
Chen *et al*. [5]	97.1	99.4	83.2	93.8	100.3	115.4	95.2	96.9	111.4	112.7	94.1	104.1	101.5	86.3	96.5	99.2
Ours(2-frame)	89.9	92.4	78.5	91.8	93.0	97.1	88.7	86.4	97.1	101.0	89.2	98.3	90.3	71.5	79.0	89.6
Ours(5-frame)	88.0	89.6	75.0	91.3	90.9	93.5	86.8	81.5	93.7	100.3	88.0	97.2	87.6	70.8	78.4	87.5
Ours(10-frame)	87.4	89.1	74.7	90.2	90.5	93.3	86.2	80.1	93.5	99.7	87.8	96.4	87.2	70.3	77.6	86.9
Ours(20-frame)	**87.1**	**88.7**	**74.6**	**90.0**	**90.3**	**93.4**	**85.7**	**80.2**	**93.1**	**99.9**	**87.4**	**96.3**	**87.2**	**70.0**	**77.1**	**86.7**

Therefore our proposed method has the advantage that it can be generalized to any natural videos.

We observed that the choice of the sampling strategy in our proposed method is crucial. Previous methods for 3D pose estimation in video always assume that the camera view is fixed. On the contrary in our MVMP model we assume that the camera view is moving as the person is moving. That is why we propose another sampling strategy which aim at making interval between two sampled frames longer. We reason that this new sampling strategy can take more advantage of temporal information. We investigate different sampling strategies during our evaluation. We implemented a naive sampling strategy that samples adjacent frames and another strategy of sampling frames with much longer interval from the input video. The naive strategy is sampling from frame number $k * n + 1$ to $k * n + k$ while the other one is sampling $n, (L/k) + n, (L * 2/k) + n, ..., (L * (k-1)/k) + n$. The second strategy outperforms the first one by about 3 mm in the condition of sampling 20 frames from a video. Moreover, with second strategy we tried sampling 2 frames, 5 frames, 10 frames, and 20 frames from the input video. Approach of sampling 20 frames outperforms approach of sampling 2 frames by 3mm. The results of different numbers of sampling are shown in Fig.4. With using only the symmetrical loss, the performance of our model is similar to the performance of [5]; with using both the symmetrical loss and the temporal loss we are able to achieve a better performance compared to state-of-the-art method [5]. We illustrate our qualitative results in Fig.5 including some visible improvements due to our body consistency constraints compared with state-of-the-art unsupervised approaches [5].

V. CONCLUSION AND FUTURE WORK

In this paper we propose our model which is extended from single-frame GAN approach to process multi-view-multi-pose(MVMP) 3D human pose estimation in videos. We implement a loose body consistency constraint relying on the symmetry within a single frame and body consistency over different frames. This constraint improves the result by about 10%. Additionally we propose a sampling strategy to sample frames from the input video and make the interval between two adjacent sampled frames as big as possible, which aims to take more advantage of temporal information in the input video. This strategy improves the result by about 3mm.

In the future we plan to implement state-of-the-art multi-view approach on our MVMP model, inspired by Iskakov *et al*. [28]. We reason that there is a way to combine the state-of-the-art multi-view 3D pose estimation approach with our model which aims to estimate 3D pose in a video.

REFERENCES

[1] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2640–2649.

[2] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, "3d human pose estimation in the wild by adversarial learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5255–5264.

[3] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7753–7762.

[4] Y. Kudo, K. Ogaki, Y. Matsui, and Y. Odagiri, "Unsupervised adversarial learning of 3d human pose from 2d joint locations," *arXiv preprint arXiv:1803.08244*, 2018.

[5] C.-H. Chen, A. Tyagi, A. Agrawal, D. Drover, S. Stojanov, and J. M. Rehg, "Unsupervised 3d pose estimation with geometric self-supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5714–5724.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[7] F. Moreno-Noguer, "3d human pose estimation from a single image via distance matrix regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2823–2832.

[8] C.-H. Chen and D. Ramanan, "3d human pose estimation= 2d pose estimation+ matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7035–7043.

[9] H. Jiang, "3d human pose reconstruction using millions of exemplars," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 1674–1677.

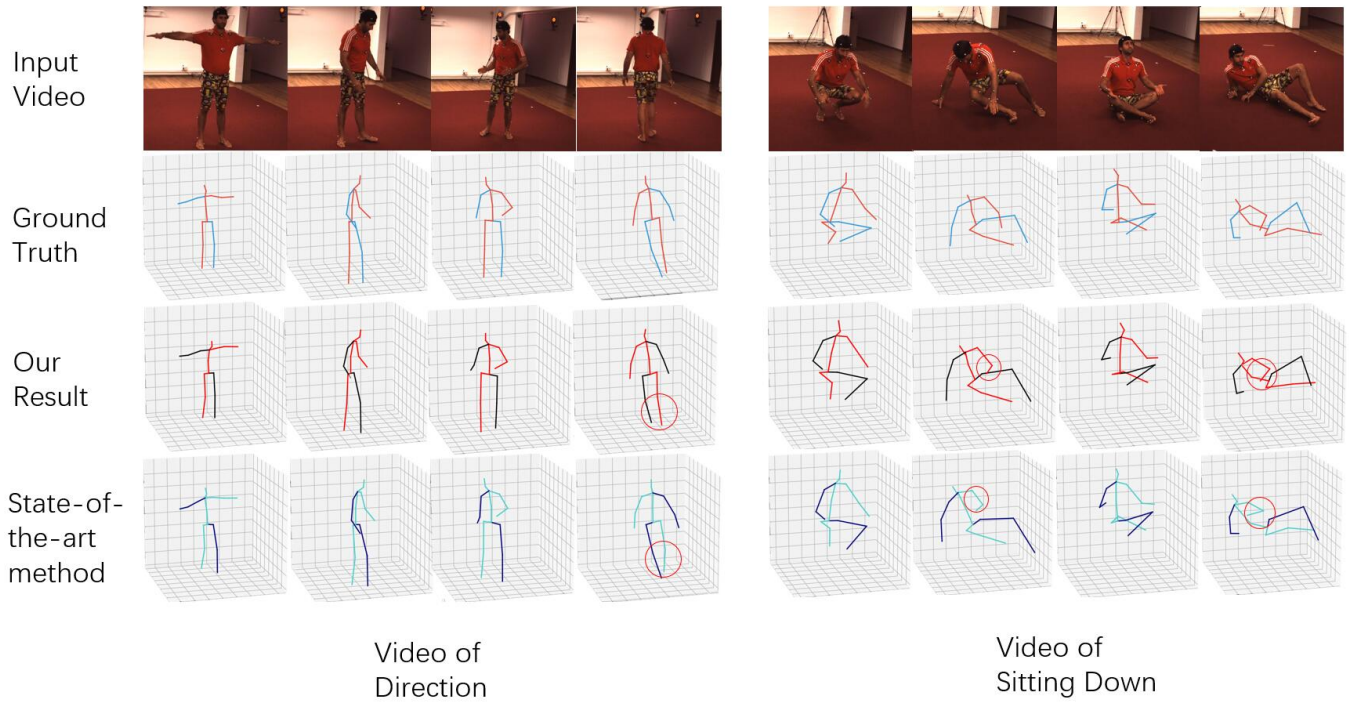

Fig. 5. Qualitative results. The first line includes input images; the second line includes 3D ground truth; the third line includes our 3D pose estimation; the fourth line includes our implementation of state-of-the-art method [5], in which we circle some improvements due to our body consistency constraints in red color

[10] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall, "A dual-source approach for 3d pose estimation from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4948–4956.

[11] B. Tekin, P. Márquez-Neila, M. Salzmann, and P. Fua, "Learning to fuse 2d and 3d image cues for monocular body pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3941–3950.

[12] M. Wang, X. Chen, W. Liu, C. Qian, L. Lin, and L. Ma, "Dr-pose3d: Depth ranking in 3d human pose estimation," *arXiv preprint arXiv:1805.08973*, 2018.

[13] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2602–2611.

[14] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis, "Sparseness meets deepness: 3d human pose estimation from monocular video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4966–4975.

[15] E. Brau and H. Jiang, "3d human pose estimation via deep learning from 2d annotations," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 582–591.

[16] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3d pose estimation from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2500–2509.

[17] H.-Y. F. Tung, A. W. Harley, W. Seto, and K. Fragkiadaki, "Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 4364–4372.

[18] M. R. Ronchi, O. Mac Aodha, R. Eng, and P. Perona, "It's all relative: Monocular 3d human pose estimation from weakly supervised data," *arXiv preprint arXiv:1805.06880*, 2018.

[19] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, A. Sharma, and A. Jain, "Learning 3d human pose from structure and motion," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 668–683.

[20] H. Rhodin, J. Spörri, I. Katircioglu, V. Constantin, F. Meyer, E. Müller, M. Salzmann, and P. Fua, "Learning monocular 3d human pose estimation from multi-view images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8437–8446.

[21] H. Rhodin, M. Salzmann, and P. Fua, "Unsupervised geometry-aware representation for 3d human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 750–767.

[22] D. Drover, C.-H. Chen, A. Agrawal, A. Tyagi, and C. Phuoc Huynh, "Can 3d pose be learned from 2d projections alone?" in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[23] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua, "Direct prediction of 3d body poses from motion compensated sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 991–1000.

[24] I. Katircioglu, B. Tekin, M. Salzmann, V. Lepetit, and P. Fua, "Learning latent representations of 3d human pose with deep neural networks," *International Journal of Computer Vision*, vol. 126, no. 12, pp. 1326–1341, 2018.

[25] M. Rayat Imtiaz Hossain and J. J. Little, "Exploiting temporal information for 3d human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 68–84.

[26] K. Lee, I. Lee, and S. Lee, "Propagating lstm: 3d pose estimation based on joint interdependency," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 119–135.

[27] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, 2017, pp. 214–223.

[28] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov, "Learnable triangulation of human pose," *arXiv preprint arXiv:1905.05754*, 2019.

[29] M. Kocabas, S. Karagoz, and E. Akbas, "Self-supervised learning of 3d human pose using multi-view geometry," *arXiv preprint arXiv:1903.02330*, 2019.

[30] B. Wandt and B. Rosenhahn, "Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7782–7791.