School of Computer and Communication sciences

3D Human Pose Estimation using a Part-Based Discriminator

Student

Mehdi Ziazi

mehdi.ziazi@epfl.ch

Supervisor

Soumava Kumar Roy

soumava.roy@epfl.ch

CVLAB

EPFL, SWITZERLAND

Abstract

3D pose estimation methods are shown to be remarkably effective, provided there exists a large corpus of abundant labeled data. However generating this large corpus of annotations is a labor intensive and time consuming task besides the high degree of expertise Therefore much of the recent focus in this field has shifted to semi/weakly/self supervised learning. Even though these methods have shown a remarkable improvement in their performance over recent years, they still suffer from occlusion, changes in illumination, low image resolution etc.. *This results in noisy key-points detection leading to 3D pose estimation. Thus in this project, we attempt to reduce the impact of noisy key-points by utilizing a part-based KCS discriminator on the estimated 3D poses by a 2D-3D lifting network. The discriminator consists of five sub-discriminators to impose pose constraints on five different regions of the predicted 3D pose, which in turn holistically imposes these constraints on the whole human body. Our experimental results on the Human 3.6M and MPI-INF-3DHP datasets have demonstrated the benefit of using the proposed 3D discriminator on the predicted 3D poses.*

Contents

List of Figures

List of Tables

1. Introduction

3D Human Pose Estimation methods aim to predict the 3D body joints so as to build a full human body representation from input data such as images and videos. It is a fundamental task that provides a geometrical and structural information of the human body, with a number of broad applications in action recognition/prediction/correction [25, 28], human-robot interaction [13], human tracking [17], motion analysis, *etc*. This task is typically solved using a large corpus of ground truth 3D annotated data which is collected in an indoor environmental setup using motion capture systems. As a result, supervised approaches to solve this task have become remarkably effective over the recent years [26, 30, 31]. However, there exists many different scenarios that involve in-the-wild motion activities which cannot be sufficiently captured by such indoor data collection setups. Thus recently much of the attention has shifted to semi/self/weakly supervised approaches which attempt to generate a satisfactory pseudo labels for the unannotated samples [1, 7, 19].

In general, these semi-supervised approaches are vulnerable to

- changes in illumination and view points.
- changes in image resolutions.
- occlusions.

This in-turn leads to noisy key-points detections and generation of noisy (inaccurate) 3D pseudo labels. Occlusion still remains one of the most profound difficulties in 3D HPE methods, and its detrimental impact is further aggravated in single view images due to the inherent depth ambiguity of the predicted 3D joints. A number of works attempt to alleviate the effect of these noisy detections by exploiting temporal consistencies between the predicted joints over a temporal sequence of images [2, 19]. Nevertheless, these result in complex prediction networks which are difficult to train and typically need other forms of auxiliary supervision (such as presence 2D GT heatmaps/joints) to achieve competitive results. Others [20, 23] impose pose priors and kinematic pose constraints on the entire human body to neutralize the impact of noisy detections. However, such methods impose these constraints on the predicted joints of the entire body as a whole, while ignoring the predictions of each part(s) of the body separately; leading to sub-optimal results.

Motivation: To overcome these limitations, we propose to make use of a 3D pose discriminator on the predicted 3D poses so as to reduce the impact of noisy predicted key-points on the estimated 3D pose. More specifically, the discriminator aims to separately detect irregular 3D joint predictions in each of the five body parts (namely *left leg*, *right leg*, *left hand*, *right hand* and *torso*). Using such part-based discriminators allows us to enforce pose-prior constraints in each of these aforementioned body parts, which in-turn holistically imposes such constraints on the entire human body.

2. Method

Our goal is to train a deep neural network so as to estimate 3D human poses from single view images in the semi-supervised setup using as little annotated data as needed. Thus, we first obtain 2D detected key-points from each single view image using the detector of [24]. This is followed by a simple MLP based *lifting network* which infers 3D poses from the detected 2D ones. The predicted 3D poses by the lifting network are fed in to the part-based KCS discriminator which attempts to impose part-based pose constraints on them to reduce the impact of noisy predicted 3D key-points.

Notations: Let $\mathcal{L} = \{\boldsymbol{I}^l, \boldsymbol{p}^l_{3D}\}_{l=1}^{\mathrm{N_L}}$ denote the set of labeled single view RGB images $\boldsymbol{I}^l \in \mathbb{R}^{h\times w\times 3}$, featuring a person. $\boldsymbol{p}^l_{3D}$ denotes the ground-truth 3D pose of the person's body and l is the image index. $\boldsymbol{p}^l_{3D}$ represents the full set of $\mathrm{N_J}$ body joints $\{\boldsymbol{X}^l_j\}_{j=1}^{\mathrm{N_J}}$, where $\boldsymbol{X}^l_j \in \mathbb{R}^3$ denotes the 3D coordinates of joint j in image l. Similarly, let $\mathcal{U} = \{\boldsymbol{I}^u\}_{u=1}^{\mathrm{N_U}}$ be a longer sequence of $\mathrm{N_U}$ images, but without associated body poses. We drop the l and u notations to avoid any clutter of notations when there is no ambiguity. The 2D-3D lifting network with learnable parameters $\boldsymbol{\Phi}$ is denoted as $f_{\boldsymbol{\Phi}}$, where as KCS part-based Discriminator with learnable parameters $\boldsymbol{\Theta}$ is denoted as $g_{\boldsymbol{\Theta}}$.

2.1. Lifting 2D poses to 3D:

We begin with the 2D joint detections obtained using the 2D pose estimator network of [24]. Thereafter, we pass these 2D detections through $f_{\boldsymbol{\Phi}}$ that maps (or lifts) these 2D detections to 3D pose $\hat{\boldsymbol{p}}_{3D}$, which are treated as estimates to the true 3D pose $\boldsymbol{p}_{3D}$. Here, we adopt a commonly used representation of [5, 6], namely 2.5D $\boldsymbol{p}^{2.5D} = \{(\boldsymbol{u}_j, \boldsymbol{v}_j, \boldsymbol{d}^{root} + \boldsymbol{d}^{rel}_j)\}_{j=1}^{\mathrm{N_J}}$ where $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ denote the components of joint j^{th} in the undistorted 2D image space, $\boldsymbol{d}^{root}$ is a scalar representing the depth of the root (or pelvis) joint with respect to the camera and $\boldsymbol{d}^{rel}_j$ is the relative depth of each joint to the root.

In order to obtain the 3D poses, we first train a MLP similar to [15] to obtain the root relative depths of every joint $d^{rel}_j = g_\phi(f_{\boldsymbol{\Phi}}(\boldsymbol{I}))$. Thereafter, the 3D pose in the camera or the world coordinate system is recovered using the inverse of the projection equations as shown below:

$$(\boldsymbol{d}^{root} + \boldsymbol{d}^{rel}_j)\begin{pmatrix}\boldsymbol{u}_j\\ \boldsymbol{v}_j\\ 1\end{pmatrix} = \boldsymbol{K}\begin{pmatrix}\boldsymbol{X}^C_j\\ \boldsymbol{Y}^C_j\\ (\boldsymbol{d}^{root} + \boldsymbol{d}^{rel}_j)\end{pmatrix} = \boldsymbol{K}\left[\boldsymbol{R}\begin{pmatrix}\boldsymbol{X}_j\\ \boldsymbol{Y}_j\\ \boldsymbol{Z}_j\end{pmatrix} + \boldsymbol{t}\right] \quad , \tag{1}$$

where $\boldsymbol{X}^C$ and $\boldsymbol{Y}^C$ denote the first two components of a joint in camera space, $(\boldsymbol{X}_j, \boldsymbol{Y}_j, \boldsymbol{Z}_j)$ is the joint in the world coordinate and $\boldsymbol{K} \in \mathbb{R}^{3\times3}$, $\boldsymbol{R} \in \mathbb{R}^{3\times3}$, and $\boldsymbol{t} \in \mathbb{R}^{3\times1}$ represent the intrinsic matrix, rotation matrix, and translation vector of the camera respectively. Eqn. (1) assumes that the depth of the root joint d^{root} is known beforehand. Similar to the works of [15, 22, 26], we set the position of the root joint to its ground truth value; however an approximation of it can be obtain analytically as shown in [5].

2.2. Part-Based KCS Discriminator:

The **K**inematic **C**hain **S**pace (KCS) [27] decomposes the human posture in terms of joint angles and bone lengths. The bone vector between the $\mathbf{r}^t h$ and $\mathbf{t}^t h$ joints is shown as below:

$$\boldsymbol{b}_{rt} = \boldsymbol{X}_r - \boldsymbol{X}_t = \boldsymbol{C}_{rt}\boldsymbol{X} \quad , \tag{2}$$

where $\boldsymbol{X} \in \mathbb{R}^{\mathrm{N_J}\times 3}$ denotes the 3D pose in world/camera coordinates and $\boldsymbol{C}_{rt}$ is a vector of size $\mathrm{N_J}$ whose elements are follows:

$$\boldsymbol{C}_{rt}^{j} = \begin{bmatrix} 1 & \text{if } j == r \\ 0 & \text{otherwise} \\ -1 & \text{if } j == t \end{bmatrix} \quad , \quad 1 \leq j \leq \mathrm{N_J} \tag{3}$$

Therefore, we concatenate each of the $\mathrm{N_B}$ bone vectors $\boldsymbol{b}_{rt} \in \mathbb{R}^3$ obtained in Eqn. (2) as columns to obtain the bone matrix $\boldsymbol{B} \in \mathbb{R}^{3\times \mathrm{N_B}}$. The final KCS Matrix $\boldsymbol{\Psi}$ is obtained as shown below:

$$\boldsymbol{\Psi} = \boldsymbol{B}^\top \boldsymbol{B}. \tag{4}$$

As a result of the inner product between two bone vectors in Eqn. (4), $\boldsymbol{\Psi}$ consists of bone lengths and a (scaled) angular representation on the diagonal and the remaining entries respectively. Therefore, a number of works have used an additional layer to generate the $\boldsymbol{\Psi}$, thus by-passing the need to compute the bone lengths and the angular constraints. Furthermore, $\boldsymbol{\Psi}$ can also be used to impose symmetric bone length constraints on the predicted 3D poses due to its inherent symmetric nature.

We design a part-based KCS matrix as the input to the 3D pose discriminator. Specifically, we divide the human body in to five different *parts*, namely left leg, right leg, left hand, right hand and torso; then generate a part-wise KCS matrix for each of them separately. Instead of analysing the 3D pose of the entire body, such part-wise implementations considers the local (or part) body bone lengths and bone angles, so as to decipher the poses of each part individually and subsequently discriminate the 3D pose of the entire human body. The part-based KCS matrix $\boldsymbol{\Psi}_i$ is defined as below

$$\boldsymbol{\Psi}_i = \boldsymbol{B}_i^\top \boldsymbol{B}_i \quad , \tag{5}$$

where i denotes either of the five parts considered and $\boldsymbol{B}_i$ denotes the corresponding bone matrix of the i^{th} part. $\boldsymbol{B}_i$ is obtained by considering only the joints belonging to the i^{th} part (as shown using different colors in Fig. 1). Finally, a 3D pose discriminator is constructed which takes as input all the part-based KCS matrices $\{\boldsymbol{\Psi}_i\}_{i=1}^{5}$ and trains a part sub-discriminator on each of the $\boldsymbol{\Psi}_i$ in order to discriminate the 3D poses of each part separately [3].

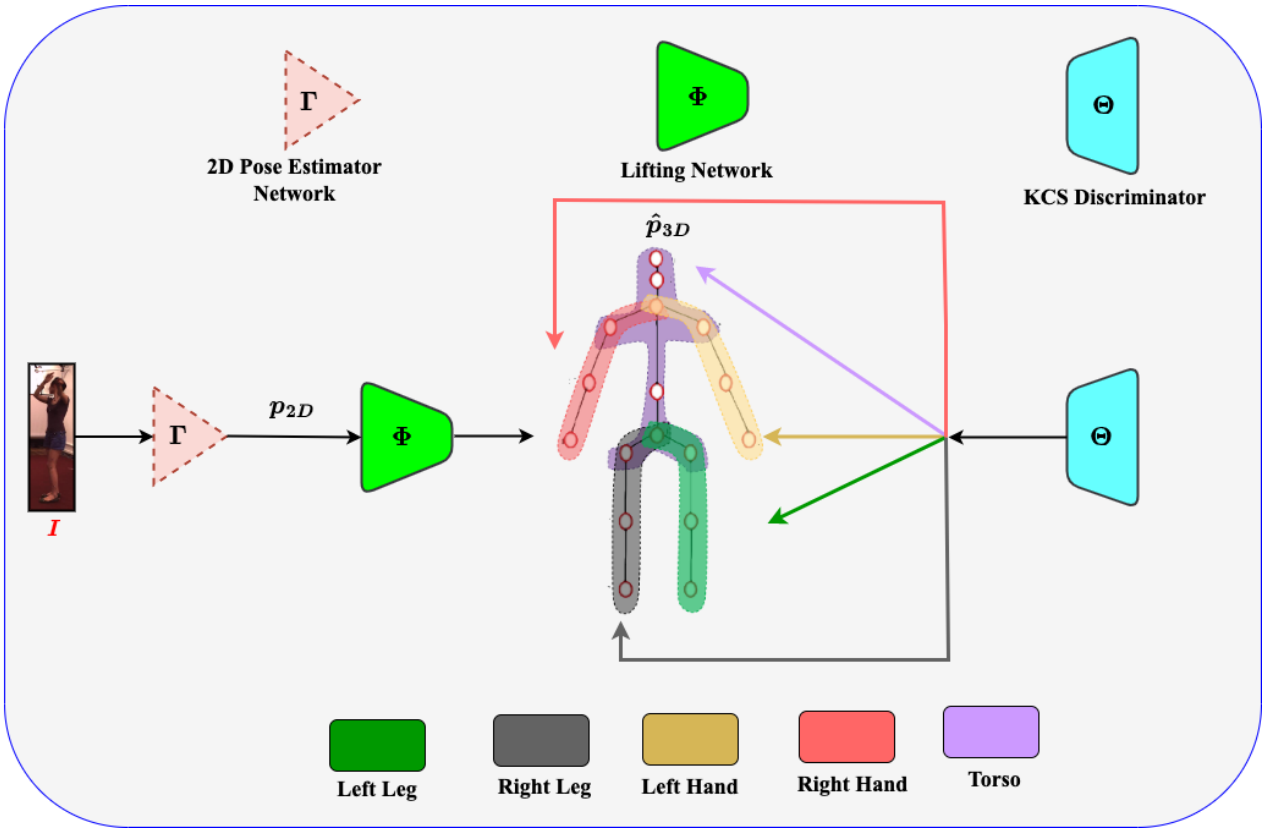

Figure 1. A schematic representation of our training setup consisting of the 2D-3D Lifting Network (with parameters $\mathbf{\Phi}$) and part-based KCS Discriminator (with parameters $\mathbf{\Theta}$). The parameters $\mathbf{\Gamma}$ of the 2D pose estimator model is kept **fixed**.

2.3. Implementation Details

Lifting Network: The 2D-3D Lifting Network $f_{\mathbf{\Theta}}$ consists of a MLP with 2 hidden layers of size 2048. The hidden layers are composed of a linear layer, each followed by ReLU activation and 10% dropout. The final layer consists of only a linear layer, with its output dimension set to $\mathrm{N_J}$ (*i.e.* the number of joints). Given the single-view input image $\boldsymbol{I}$, we pass it through the 2D pose estimator model of [24] and obtain the 2D key-points detections $\hat{\boldsymbol{p}}_{2D}$. $\hat{\boldsymbol{p}}_{2D}$ is further normalized in the range of $[-1, 1]$ and passed through $f_{\mathbf{\Phi}}$ to obtain the root relative depth $\boldsymbol{d}^{rel}$ of every joint.

Discriminator Network: The part-based KCS discriminator consists of 4 hidden linear layers with leaky ReLU and residual connections, where the hidden feature dimensions are set to 1000. A schematic of discriminator functioning on the "*torso*" is shown in Fig. 2.

Training Losses: The loss function to train the parameters $\mathbf{\Phi}$ and $\mathbf{\Theta}$ is given below:

$$L_{\text{Train}}(\mathbf{\Phi}, \mathbf{\Theta}) = L_{\text{lift}} + \lambda_{\text{dis}}\, L_{\text{dis}} \quad . \tag{6}$$

The lifting loss L_{lift} is as follows:

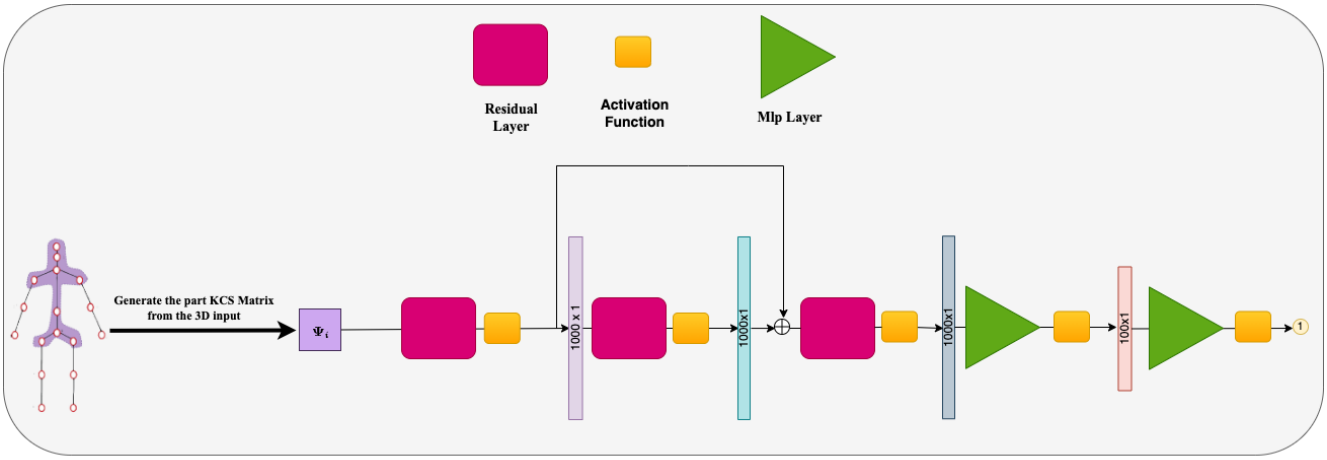

Figure 2. A schematic representation of the part-based KCS discriminator operating on the *torso* region. Similar part-based KCS discriminators are used for the other four different parts of the human body.

$$L_{\text{lift}} = L_{\text{MSE}}\left(\hat{\boldsymbol{p}}_{3D}, \boldsymbol{p}^t_{3D}\right) \ , \tag{7}$$

where L_{MSE} is mean squared error loss, while $\hat{\boldsymbol{p}}_{3D}$ denotes the 3D pose predicted by the lifting network $f_{\boldsymbol{\Theta}}$ and $\boldsymbol{p}^t_{3D}$ denotes the target 3D pose. Further, we adopt the well known LS-GAN loss [14] as the discriminator loss L_{dis} which is shown below:

$$L_{\text{dis}} = \mathbb{E}\left[g_{\boldsymbol{\Theta}}(\boldsymbol{p}^t_{3D}) - 1\right] + \mathbb{E}\left[g_{\boldsymbol{\Theta}}(\hat{\boldsymbol{p}}_{3D})\right] \tag{8}$$

In Eqn. (7) and (8), the target 3d pose $\boldsymbol{p}^t_{3D}$ is set to the ground truth 3D poses for the samples belonging to the labeled set $\mathcal{L}$; while for the samples belonging to the unlabeled set $\mathcal{U}$, it is set to the pseudo 3D poses generated generated by running the robust weighted triangulation scheme of [24] on the joints detected by the 2D pose estimator $h_{\boldsymbol{\Gamma}}$. We report the results after training both the networks for 100 epochs. We have used the Adam optimizer [8] where the learning rate of $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ are set 1×10^{-4} and 5×10^{-4} respectively. The training batch size is set to 256. The value of λ_{dis} is set to 0.5 in Eqn. (6) for all the experiments.

3. Experiments

We evaluate the functioning of the 3D pose discriminator in the semi-supervised setting on two well known datasets (**a**) Human 3.6M [4] and (**b**) MPI-INF-3DHP [16].

3.1. Datasets and Metrics

In this section, we provide a brief description of the datasets and the evaluation metrics used in our experiments.

Table 1. Quantitative results Human3.6M (Semi-supervised). Best results are shown in **bold**.

10% of All Data			
Method	**MPJPE** ↓	**NMPJPE** ↓	**PMPJPE** ↓
Kundu *et.al.* [10]	-	-	50.8
Roy *et al*. [24]	56.9	56.6	45.4
Ours	**56.1**	**56.0**	**44.7**
Only S1			
Rhodin *et al*. [21]	131.7	122.6	98.2
Pavlako *et al*. [18]	110.7	97.6	74.5
Li *et al*. [12]	88.8	80.1	66.5
Rhodin *et al*. [22]	-	80.1	65.1
Kocabas *et al*. [9]	-	67.0	60.2
Pavllo *et al*. [19]	64.7	61.8	-
Iqbal *et al*. [6]	62.8	59.6	51.4
Kundu *et al*. [10]	-	-	52
Roy *et al*. [24]	60.8	60.4	48.4
Ours	**60.2**	**60**	**48.1**

Human 3.6M [4]: It is the most widely used indoor dataset for 3D pose estimation using single or multiple cameras. The dataset comprises of 3.6 million images that are captured using 4 different but calibrated cameras. As in most published papers, we use the subjects S1, S5, S6, S7, S8 in the training phase, while the subjects S9, S11 are used to evaluate the performance of our 2D-3D lifting network $LNet_{\mathbf{\Phi}}$. In the semi-supervised learning setup, we evaluate our method on two setups as follows:

- similar to the training protocols of [11, 18, 22], the supervised set $\mathcal{L}$ consists of images of the training subject S1, while the remaining subjects constitute the unsupervised set $\mathcal{U}$.
- The supervised set $\mathcal{L}$ consist of 10% of the training samples, while the remaining samples make up $\mathcal{U}$.

MPI-INF-3DHP [16]: This dataset consists of both indoor and outdoor images for single person 3D pose estimation. There are 8 subjects performing different actions which are captured using 8 different calibrated cameras, thereby covering a wide range of 3D poses and view angles. It contains both indoor and complex outdoor scenes featuring a wide range of actions ranging from walking, sitting to challenging exercise poses with dynamic motions that are difficult to estimate. In the semi-supervised setup, we use the images belonging to subject S1 as the labeled set $\mathcal{L}$ and the other subjects form the unlabeled set $\mathcal{U}$. We also only use the 5 chest cameras in the training of our models.

Metrics. During the training phase, we perform end-to-end training and learn the parameters of the whole network as shown in Fig. 1. This results in a fully trained lifting network $LNet_{\mathbf{\Phi}}$ that can operate on single view images. Hence, we report the Mean Per-Joint Position Error (**MPJPE**), the normalized **NMPJPE**, and the procrustes aligned **PMPJPE** in millimeters between the predictions of the 2D-3D lifting network $LNet_{\mathbf{\Phi}}$ and the ground truth 3D poses on the test set.

Table 2. Quantitative results MPI-INF-3DHP (Semi-supervised). Best results are shown in **bold**.

Method	MPJPE ↓	NMPJPE ↓	PMPJPE ↓
Rhodin *et al*. [22]	-	121.8	-
Kocabas *et al*. [9]	-	119.9	-
Iqbal *et al*. [6]	113.8	102.2	-
Roy *et al*. [24]	102.2	99.6	93.6
Ours	**101.7**	**99.5**	**92.7**

Note: We train and evaluate our models on every 5^{th} frame on both these datasets.

3.2. Quantitative Results - Semi Supervised

Table 1 shows the results of using the part-based KCS discriminator on the Human 3.6M [4] dataset. Considering only 10% of the data as labeled, we outperform both the competing method (*i.e.* Kundu *et.al.* [10] and Roy *et al*. [24]) across all the three evaluation metrics. Likewise, we also perform better than the other competing methods when considering all the samples of S1 as labeled (Only S1). Method such as Kundu *et al*. [10] rely on additional datasets such as in-the-wild YouTube videos and the MADS [29], in addition to a part-based puppet to instill prior human skeletal knowledge in their learning framework; while we do not. This clearly indicates that the part-based KCS discriminator is able to better reason about the unwanted poses occurring in various parts of the predicted human pose, thereby leading to superior results over the other methods.

The results on the MPI-INF-3DHP [16] dataset is shown in Table 2, where we again outperform the other baseline methods in terms of the evaluation metrics by just using the proposed part-based KCS discriminator, without having to use any additional training datasets.

4. Conclusion

In this work, we have successfully integrated a part based KCS discriminator in the framework of predicting 3D poses from single view images. The discriminator creates a kinematic chain space for each of the five distinct body parts separately, and aims to detect irregular 3D predictions so as to impose pose-prior constraints holistically on the entire human body. The quantitative experimental results obtained on two well known datasets clearly demonstrate the effectiveness of the discriminator used in the learning framework (Refer to Table 1 and 2 on Section § 3.2). In the future, we aim to expand the functionality of the part based KCS discriminator in order to enforce temporal motion consistency while learning from a temporal sequence in monocular videos. At present, the performance of our method is limited by the detections of the 2D pose estimator model of Roy *et al*. [24]; which we will attempt to improve by imposing pose priors on the triangulated 3D poses. Similar to Kundu *et al*. [10], we can also predict the camera parameters, which can circumvent the need of using well calibrated cameras in the learning framework in the future.

References

[1] C. Chen, A. Tyagi, A. Agrawal, D. Drover, R. MV, S. Stojanov, and J. M. Rehg. Unsupervised 3D Pose Estimation with Geometric Selfsupervision. In *CVPR*, 2019.

[2] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. Occlusion-Aware Networks for 3D Human Pose Estimation in Video. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 723–732, 2019.

[3] Kehong Gong, Jianfeng Zhang, and Jiashi Feng. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8575–8584, June 2021.

[4] C. Ionescu, I. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE TPAMI*, 2014.

[5] U. Iqbal, P. Molchanov, T. Breuel Juergen Gall, and J. Kautz. Hand Pose Estimation via Latent 2.5 D Heatmap Regression. In *ECCV*, pages 118–134, 2018.

[6] U. Iqbal, P. Molchanov, and J. Kautz. Weakly-Supervised 3D Human Pose Learning via Multi-View Images in the Wild. In *CVPR*, 2020.

[7] A. Kanazawa, J. Zhang, P. Felsen, and J. Malik. Learning 3D Human Dynamics from Video. In *CVPR*, 2019.

[8] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimisation. In *ICLR*, 2015.

[9] M. Kocabas, S. Karagoz, and E. Akbas. Self-Supervised Learning of 3D Human Pose Using Multi-View Geometry. In *CVPR*, 2019.

[10] J. N. Kundu, S. Seth, V. Jampani, M. Rakesh, R. V. Babu, and A. Chakraborty. Self-Supervised 3D Human Pose Estimation via Part Guided Novel Image Synthesis. In *CVPR*, 2020.

[11] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu. High Performance Visual Tracking with Siamese Region Proposal Network. In *CVPR*, June 2018.

[12] Z. Li, X. Wang, F. Wang, and P. Jiang. On Boosting Single-Frame 3D Human Pose Estimation via Monocular Videos. In *ICCV*, 2019.

[13] I Scott MacKenzie. Human-Computer Interaction: An Empirical Research Perspective. 2012.

[14] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least Squares Generative Adversarial Networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[15] J. Martinez, R. Hossain, J. Romero, and J.J. Little. A Simple Yet Effective Baseline for 3D Human Pose Estimation. In *ICCV*, 2017.

[16] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. 2017.

[17] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-Time 3D Human Pose Estimation with a Single RGB Camera. *Acm transactions on graphics (tog)*, 36(4):1–14, 2017.

[18] G. Pavlakos, N. Kolotouros, and K. Daniilidis. Texturepose: Supervising Human Mesh Estimation with Texture Consistency. In *ICCV*, 2019.

[19] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. 3D Human Pose Estimation in Video with Temporal Convolutions and Semi-Supervised Training. In *CVPR*, 2019.

[20] Ibrahim Radwan, Abhinav Dhall, and Roland Goecke. Monocular Image 3D Human Pose Estimation under Self-Occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1888–1895, 2013.

[21] H. Rhodin, M. Salzmann, and P. Fua. Unsupervised Geometry-Aware Representation for 3D Human Pose Estimation. In *ECCV*, 2018.

[22] H. Rhodin, J. Spoerri, I. Katircioglu, V. Constantin, F. Meyer, E. Moeller, M. Salzmann, and P. Fua. Learning Monocular 3D Human Pose Estimation from Multi-View Images. In *CVPR*, 2018.

[23] G. Rogez, P. Weinzaepfel, and C. Schmid. Lcr-Net: Localization-Classification-Regression for Human Pose. In *CVPR*, 2017.

[24] S. Roy, L. Citraro, S. Honari, and P. Fua. On Triangulation as a Form of Self-Supervision for 3D Human Pose Estimation. 2022.

[25] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1227–1236, 2019.

[26] X. Sun, F. Wei B. Xiao, S. Liang, and Y. Wei. Integral Human Pose Regression. In *ECCV*, 2018.

[27] Bastian Wandt and Bodo Rosenhahn. 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 (CVPR)*, June 2019.

[28] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[29] W. Zhang, Z. Liu, L. Zhou, H. Leung, and A. Chan. Martial Arts, Dancing and Sports Dataset: A Challenging Stereo and Multi-View Dataset for 3D Human Pose Estimation. 61:22–39, 2017.

[30] X. Zhou, Q. Huang, X. Sun, X. Xue, and A. Y. Wei. Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Approach. In *ICCV*, 2017.

[31] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep Kinematic Pose Regression. In *ECCV*, 2016.