Monocular 3D Pose Estimation with Uncertainty Estimation for Handling Occlusion.

Research Project Proposal

CVLAB, EPFL

1. Description

Supervised methods have been quite successful in the recent years for estimating 3D human poses from single view images, provided that enough 3D annotated data is available [12, 16–18]. However such methods usually fail in in-the-wild images captured from a wide number of scenarios involving unusual activities, for which acquiring the annotations become a challenging task. Thus, much of the recent focus has shifted to Semi/Weakly/Self supervised methods [1, 7, 8, 11], which are aimed to provide a reliable pseudo-label estimate for the unannotated single view images.

However, the Semi/Weakly/Self supervised methods are highly vulnerable to (**a**) occlusion, (**b**) changes in illumination and viewing angles, (**c**) low image resolution *etc*., thereby resulting in noisy 2D estimates and 3D pseudo targets; out of which **occlusion** remains the primary and a major source of such noisy estimates. A number of works aim to tackle occlusions using multi-view camera setup where the pseudo targets are obtained using multi-view geometry of well calibrated cameras [6, 9, 15]. However, they do not generalize well to single, potentially moving, cameras. A number of approaches also attempt to overcome these limitations by exploiting temporal consistency in monocular videos [2, 11]. Some works also impose prior pose and kinematic priors on the estimated 3D poses to handle such noisy detections as an additional module [13, 14]. However, the aforementioned algorithms do not explicitly model occlusion, in order to enhance the robustness of the underlying networks against it in their learning framework.

Motivation: Therefore, in this project our major goal is to increase the robustness of our prediction networks against occlusion for 3D human pose estimation. In order to achieve the above stated goal, we model occlusion for every joint as an aleatoric uncertainty estimate measure [3, 10]. More specifically, we make use of an "Uncertainty Estimation Network" which outputs the *mean* and the *standard deviation* for each of the predicted joints; which can further used to estimate the uncertainty of every joint as a uni/multivariate gaussian distribution. As a result, we can obtain a measure of uncertainty in the distribution of the 3D joints obtained from the teacher network $\text{Net}_{\mathbf{\Gamma}}$.

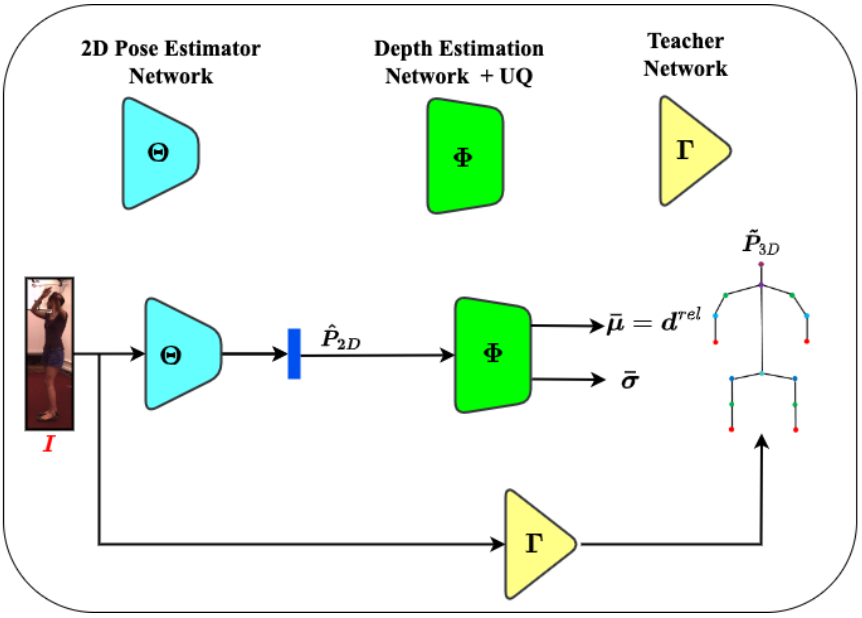

Figure 1. A schematic of the proposed method to estimate uncertainty estimates (*i.e.*$\bar{\mu}$ and $\bar{\sigma}$) for every estimated 3D joints. Here, we model only the uncertainty on depth estimates as 1D Gaussian distribution.

2. Deep Neural Networks used in the projects:

We adopt a commonly used pose representation [4, 5] namely 2.5D $\boldsymbol{p}^{2.5D} = \{(\boldsymbol{u}_j, \boldsymbol{v}_j, \boldsymbol{d}^{root} + \boldsymbol{d}_j^{rel})\}_{j=1}^{N_J}$ where $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ are the components of joint jth 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 $\boldsymbol{d}_j^{rel}$ is the relative depth of each joint to the root. The main source uncertainty lies in the estimated value of $\boldsymbol{d}_j^{rel}$ as there are infinite values of $\boldsymbol{d}_j^{rel}$ that can result in the same 2D projection on the image $\boldsymbol{I}$. We attempt to model this uncertainty estimation as a uni-variate gaussian distribution using the predictions of the "Uncertainty Estimation Network" as shown in Fig. 1. The various networks involved in the training are as follows:

1. A "2D Pose Estimator Network" with parameters $\boldsymbol{\Theta}$ which obtains the 2D joint estimates from the single view image $\boldsymbol{I}$.

2. A "Depth Estimation + **U**ncertainty **Q**uantification Network" with parameters $\boldsymbol{\Phi}$ that outputs the root relative depth $\boldsymbol{d}^{rel}$ and the standard deviation (*i.e.* $\bar{\boldsymbol{\sigma}}$) of the uni-variate gaussian distribution per joint (Refer to Fig. 1 for more details).

3. A "Teacher network" with parameters $\boldsymbol{\Gamma}$ that provides the desired pseudo labels for the unannotated samples. The weights of this network is kept fixed in general.

The loss function to train the network for uncertainty estimation only for the values of the root relative depth d_j^{rel} is shown

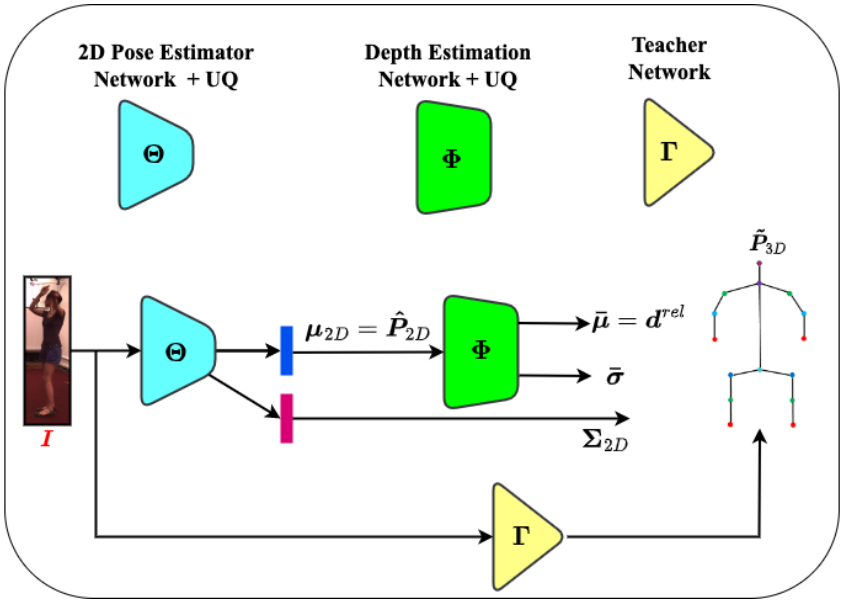

Figure 2. A schematic representation where we model the uncertainty on the 2D estimates obtained by the 2D pose estimator model along with the root relative depth. This results in a 3D multivariate gaussian distribution representing the uncertainty per joint.

below:

$$L_{Train}(\boldsymbol{\Theta}, \boldsymbol{\Phi}) = \frac{1}{\mathrm{J}} \sum_{j=1}^{\mathrm{J}} \left(\ln(\bar{\boldsymbol{\sigma}}_{\boldsymbol{j}}) + \frac{(\boldsymbol{y}_j^{rel} - \bar{\boldsymbol{\mu}}_{\boldsymbol{j}})}{\bar{\boldsymbol{\sigma}}_{\boldsymbol{j}}{}^2} \right) \quad , \tag{1}$$

where J denote the total number of joints considered in our training setup and $\boldsymbol{y}^{rel}$ denotes the target root relative depth for every joint [1].

3. Tackling the problem step by step:

1. In the first phase of the project, we will fix the value of $\bar{\boldsymbol{\sigma}}$ to a constant, and only predict the value of $\bar{\boldsymbol{\mu}}$ for every joint to model the depth-uncertainty. Here the impact of the first term in Eqn.(1) will be nullified as $\bar{\boldsymbol{\sigma}}$ is a constant value. This model will be a simple baseline which should provide some valuable insights in estimating joint uncertainties for 3D pose estimation.

2. Thereafter, we can also learn the value $\bar{\boldsymbol{\sigma}}$ along with $\bar{\boldsymbol{\mu}}$ for a comprehensive modeling of the depth uncertainty. A diagrammatic representation of this framework is shown here in Fig. 2.

3. The proposed work can further be extended by considering the uncertainty on the 2D joints obtained the 2D pose estimator network (*i.e.* u_j, v_j). This will result in a 3D multivariate gaussian which holistically attempts to learn the uncertainty in u_j, v_j and d_j^{rel}.

4. 3D Pose discriminators can also be integrated as an additional module in the learning framework to handle false positive

[1] For the un-annotated samples, $\boldsymbol{y}^{rel}$ is obtained using the Teacher network $\mathrm{Net}_{\boldsymbol{\Gamma}}$.

cases [2].

4. Prerequisites

- This project is offered to Master's students for their semester/thesis project.
- The candidate should have Python programming experience.
- Previous experience with deep learning and PyTorch, is recommended.
- Knowledge of 3D Human Pose Estimation and Uncertainty Quantification is a plus.

Contact:

Soumava Kumar Roy, EPFL Switzerland (soumava.roy@epfl.ch).

Matthias Rottmann, University of Wuppertal, Germany (rottmann@math.uni-wuppertal.de).

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. 1

[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. 1

[3] Eyke Hüllermeier and Willem Waegeman. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Machine Learning*, 110(3):457–506, 2021. 1

[4] 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. 2

[5] U. Iqbal, P. Molchanov, and J. Kautz. Weakly-Supervised 3D Human Pose Learning via Multi-View Images in the Wild. In *CVPR*, 2020. 2

[6] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov. Learnable Triangulation of Human Pose. In *ICCV*, pages 7718–7727, 2019. 1

[7] A. Kanazawa, J. Zhang, P. Felsen, and J. Malik. Learning 3D Human Dynamics from Video. In *CVPR*, 2019. 1

[8] Z. Li, X. Wang, F. Wang, and P. Jiang. On Boosting Single-Frame 3D Human Pose Estimation via Monocular Videos. In *ICCV*, 2019. 1

[9] H. Ma, L. Chen, D. Kong, Z. Wang, X. Liu, H. Tang, X. Yan, Y. Xie, S.-Y. Lin, and X. Xie. Transfusion: Cross-View Fusion with Transformer for 3D Human Pose Estimation. In *BMVC*, 2021. 1

[10] J Martin and C Elster. Aleatoric Uncertainty for Errors-in-Variables Models in Deep Regression. *Neural Processing Letters*, pages 1–20, 2022. 1

[2] A joint in $\hat{\boldsymbol{P}}_{3D}$ is a false positive if its measure of uncertainty is low, provided its prediction is incorrect.

[11] 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. 1

[12] A-.I. Popa, M. Zanfir, and C. Sminchisescu. Deep Multitask Architecture for Integrated 2D and 3D Human Sensing. In *CVPR*, 2017. 1

[13] 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. 1

[14] G. Rogez, P. Weinzaepfel, and C. Schmid. Lcr-Net: Localization-Classification-Regression for Human Pose. In *CVPR*, 2017. 1

[15] S. Roy, L. Citraro, S. Honari, and P. Fua. On Triangulation as a Form of Self-Supervision for 3D Human Pose Estimation. 2022. 1

[16] X. Sun, F. Wei B. Xiao, S. Liang, and Y. Wei. Integral Human Pose Regression. In *ECCV*, 2018. 1

[17] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua. Structured Prediction of 3D Human Pose with Deep Neural Networks. In *BMVC*, 2016. 1

[18] 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. 1