PLIKS: A Pseudo-Linear Inverse Kinematic Solver for 3D Human Body Estimation
Supplementary Material

1. Introduction

In this material, we provide implementation details and analysis of focal lengths and regularizers for our method. We further discuss the benefits of using a solver for human pose estimation utilizing constraints. Additionally, we present more qualitative results, to show the performance of PLIKS and to explore its failure scenarios.

1.1. Datasets

COCO: COCO [12] is a large-scale in-the-wild 2D keypoint dataset. We use this for training. We make use of pseudo-ground truth SMPL annotations provided by EFT [5].
MPI-INF-3DHP: MPI-INF-3DHP is an indoor multi-view and outdoor scene dataset for 3D human pose estimation. We make use of SMPL multi-view fits by SPIN [10]. We use this for training and evaluation.
Human3.6M: Human3.6M [3] is an indoor, multi-view 3D human pose estimation dataset. We follow the standard practice [6, 10] where subjects S1, S5, S6, S7, and S8 are used for training while S9 and S11 are the test subjects. We follow Protocol 2 using only the front-facing cameras.
3DPW: 3DPW [20] is a challenging outdoor benchmark for 3D pose and shape estimation. To get a fair comparison with previous state-of-the-art [8, 11], we use 3DPW training data for 3DPW experiments. We make use of a subset of this dataset PW3D-OCC following [8] for the occlusion benchmark.
AGORA: AGORA [17] is a synthetic dataset with accurate SMPL models fitted to 3D scans. The test set is not publicly available, here the evaluation is performed on the official platform. For both training and testing, we use the images of resolution 1280×720.
3DOH: 3DOH [21] is an object-occluded dataset. We use this to train and evaluate only for occlusion benchmark.
MuPoTs-3D: MuPoTs-3D [14] is a mixed indoor and outdoor multi-person dataset consisting of 20 sequences showing people performing various actions and interactions. We use this for evaluating the absolute root error.

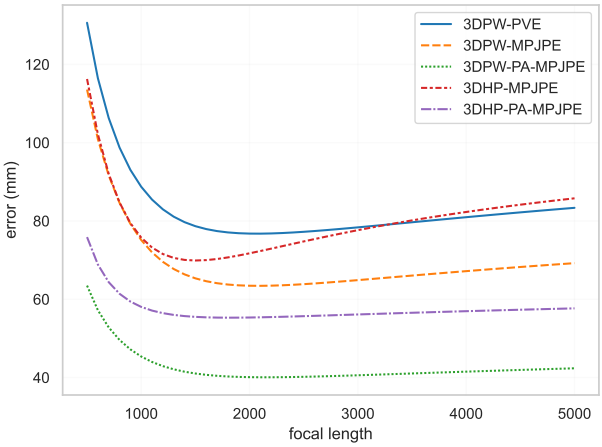

Figure 1. Impact on focal length on estimation errors when using the 3DPW [20] and MPI-INF-3DHP [13] dataset.

1.2. Network Training

As the entire pipeline is differentiable, the network is trained end-to-end. We split the training into two steps, pre-training (ARE) and training (PLIKS) to accelerate the network training speed. In pre-training, we train exclusively with the ARE module, and optimize only with respect to the mesh and network predicted parameters $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}}_k)$ by minimizing,

$$L = \omega_1 L_\theta + \omega_2 L_\beta + \omega_3 L_{2d} + \omega_4 L_{3d} + \omega_5 L_M. \quad (1)$$

Following previous work [6, 10, 15], we employ standard mesh losses to supervise the training process. Here, L_θ is the L2 loss between the predicted pose and ground truth (GT) pose. Similarly, L_β is the L2 loss between the predicted shape and GT shape. L_{2d}, L_{3d} and L_M are the L1 loss between predictions and GT 2D joint re-projection, 3D joints and, the mesh vertex in image space respectively. To supervise the 2D annotations, the predicted 3D joints are projected by the weak-perspective camera $\tilde{\boldsymbol{c}}$ as predicted by the network.

During training we make use of the PLIKS module. Due to the presence of the linear solver in PLIKS, we observe numerical instability in the early stages of training, i.e. the pixel-aligned vertex predictions are not adequately consistent for the solver, making the reconstruction ill-posed. To keep the error within bounds, we add strong shape and pose regularizers for two epochs. In this stabilization period, the shape regularizer ω_β exponentially decays from 1 to 0.1. We further add a pose-constraint to the objective function of PLIKS (Eq. (2)), such that $\omega_\theta \sum |\mathbf{\Delta R}_k| \approx \mathbf{I}$. As a consequence, the additional rotation $\mathbf{\Delta R}_k$ obtained during the stabilization period is constrained to be close to zero. Similar to ω_β, we decay ω_θ from 1 to 0. For training, we use the same objective function from Eq. (1) to minimize the mesh and the analytically predicted parameters $(\boldsymbol{\beta}, \boldsymbol{\theta}_k)$.

$$\underset{\mathbf{\Delta R}_k, \boldsymbol{\beta}, \boldsymbol{t}_k}{\operatorname{argmin}} ||\boldsymbol{w}^k \Big(\boldsymbol{i}^k - \hat{\mathbf{K}}(\mathbf{\Delta R}_k \boldsymbol{x}_r^k + \boldsymbol{\beta} \mathbf{B}_r^k + \boldsymbol{t}_k \mathbf{W}_r^k) \Big) ||_2 + \omega_\beta ||\boldsymbol{\beta}||_2. \tag{2}$$

1.3. Implementation Details

PyTorch [16] is used for implementation. For all our experiments we initialize the HRNet [18] backbone with weights pre-trained on the MPII [1] dataset, which exhibits faster convergence during training. We use the Adam optimizer [7] with a mini-batch size of 32. The learning rate at pre-training is set to $1e^{-4}$, whereas, while training the entire pipeline it is initialized to $5e^{-5}$. The network is pre-trained for 20 epochs, stabilized for 2 epochs, and then finally trained for further 30 epochs. We set the learning rate to $1e^{-5}$ while fine-tuning with the 3DPW [20] or AGORA [17] dataset. For fine-tuning, we use the previous pre-trained network as the starting point. This is to accelerate convergence and correct the 3D inaccuracies from the pseudo-GT labels. It takes around 3-5 days to train on a single NVIDIA Tesla V-100-16GB GPU. We set ω_1, ω_2, ω_3, ω_4, and ω_5 to 1, 0.05, 4, 8, and 4, respectively. As the pseudo-GT labels from EFT [5], are defined with respect to weak-perspective projection, we reduce ω_1, ω_2, ω_4, and ω_5 by a factor of 0.1 for the 2D dataset.

2. Ablations

Here we discuss the effects of shape regularizer and effects of focal length estimation.

2.1. Regularizer

To demonstrate the importance of a strong regularizer, we perform a similar experiment (from Sec 4.1) where we add random noise to the GT of the mesh vertices from the 3DPW [20] test set. Here we vary the shape regularizer weights ω_β and observe the final MPJPE obtained. From Table 1, it is evident that larger weights for ω_β is more robust to noise. However, training the network using larger weights has its own drawbacks as shown in Figure 2. The network forces the shape components $\boldsymbol{\beta}$ to always be close to zero. As the shape $\boldsymbol{\beta}$ is determined by a solver, it enables us to switch to a male, female, or neutral model seamlessly by replacing the shape coefficients $\mathbf{B}$ during inference. For our training, we set $\omega_\beta = 0.1$, as this is a good mixture between stability and shape variations.

2.2. Focal Length

We conduct experiments on the 3DPW and MPI-INF-3DHP test sets by varying the focal lengths. As shown in Fig. 1, PLIKS is robust to a wide range of focal lengths when the FOV is small (e.g., 3DPW), but it suffers from the effects of perspective warping on large focal lenghts for wide FOV images (e.g., MPI-INF-3DHP). Using CamCalib [9] on the MPI-INF-3DHP to determine the FOV and consequently the focal length of the image, we could only obtain a reduction in MPJPE of 72.01 mm, i.e., a drop of just 3%. In particular, when, the ground truth camera matrix is known, our approach can be expected to yield optimal performance.

3. Qualitative Results

In this section, we show comparisons to SOTA methods on AGORA and provide more qualitative results on various other datasets.

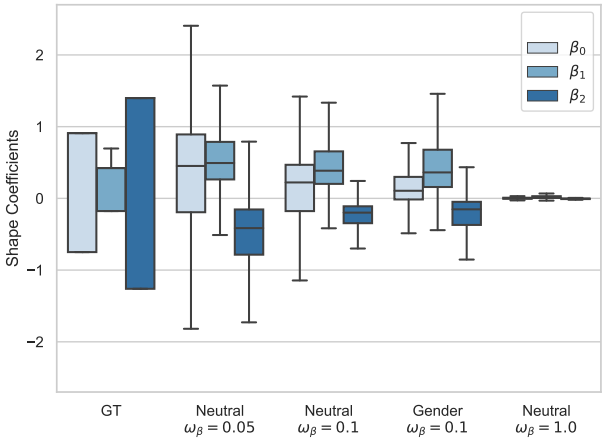

Figure 2. Effect of the regularizer weight used during training on the shape coefficients. Picking a higher ω_β reduces the error, but causes the network to output meshes getting progressively closer to the identity representation. Here neutral represents the neutral SMPL model, and gender refers to the gender-specific model on the 3DPW [20] dataset.

Figure 3. Qualitative results from AGORA test set.

	±10 mm	±20 mm	±30 mm
$\omega_\beta = 2.0$	17.3	22.6	34.9
$\omega_\beta = 1.0$	18.4	37.3	64.0
$\omega_\beta = 0.1$	122	254	322.1

Table 1. Ground truth errors in the presence of per vertex noise ranging from $\pm 10mm$ to $\pm 30mm$ and the effect of using a shape regularizer, ω_β.

3.1. Qualitative Comparison

We display several examples of PLIKS on the AGORA test set in Fig. 3. We use YOLO [4] for the bounding box estimation and CamCalib [9] for the focal length estimation. The images demonstrate that PLIKS performs better than previous approaches, by aligning the bodies well in 3D as well as 2D.

3.2. Inference Modification

As mentioned in the main paper, one of the strengths of our method is the application of constraints during inference. Here, we discuss a proof-of-concept for two use cases, where we show the benefits of using a solver without any retraining of the network. We discuss dynamic shape and translation constraints.

Dynamic Shape Although our network was trained only on a neutral SMPL model with 10 shape components, it can make use of other shape models during inference if they follow the same design principle as SMPL. As an extreme scenario, we show the application using the kid-SMPL model [17, 19]. The kid-SMPL is an extended version of the SMPL model supporting children by linearly blending the SMPL and Skinned Multi-Infant Linear Model

Figure 4. Example images with dynamic shape during inference. Set of input images, overlay and, 3D view.

(SMIL) [2] by a weighting factor $\alpha \in (0, 1)$ [17]. Here, larger weights represents infants, while smaller weights are associated with adults. For simplicity, we denote the kid-SMPL model as having 11 shape components.

Qualitative results of using the kid-SMPL model on the Relative Human (RH) dataset [19] are shown in Fig. 4. The only modification performed was adapting the shape coefficients $\mathbf{B}_r^k$ in Eq. 2 from the SMPL to their kid-SMPL counterparts. In that context, we further empirically set ω_β to 0.5. From the RH dataset we employ the GT age classifier, i.e., we use SMPL for adults, and kid-SMPL for child or infant. We observe visually satisfactory results, with sufficiently reliable depth reasoning. A top-down approach [19] or a simple age classifier could be designed to determine the age as a future work.

Translation Constraints Previous examples of just using dynamic shapes is not a complete solution, due to the ill-posed nature of the problem. This is quite evident from the fifth column of Fig. 5. As a proof-of-concept, we show the application of translation constraints during inference. We add a simple depth constraint to Eq. 2 as $\omega_t \boldsymbol{t}_{0,z}^k = \omega_t \boldsymbol{t}_{0,z}^a$. Here, $\boldsymbol{t}_{0,z}^a$ is the root depth of the adult in the image, and $\boldsymbol{t}_{0,z}^k$ is the constrained setting for the root-depth of the kids in the image, with ω_t being a weighting factor. We make the assumption that the children in the images are standing close to the adults. The solver optimizes the shape such that the translation constraint is satisfied. We empirically set ω_t to 0.2. Though, strictly not comparable, we visualize the results of BEV [19] in Fig. 5. There, all images are in fact from the RH training set on which BEV was trained.

We quickly add that this is not a real-world solution to the problem, but it emphasizes the importance of using constraints during inference or training. As future work, one could make use of the RH dataset with the depth-level information by adding a top-down approach [19] for better constraints.

3.3. Failure Mode

In Fig. 6, we show a few examples where PLIKS fails to reconstruct reasonable human body poses. The failure cases range from (a) too many people in the crop, (b) extreme poses not seen in training, and (c) extreme occlusion.

Figure 5. Example images with translation constraints during inference. (a) Input Image, (b,c) 3D overlay from BEV [19] and PLIKS respectively, (d) 3D view of the model from BEV [19], (e,f) 3D view of the model from PLIKS without and with using the translation constraint.

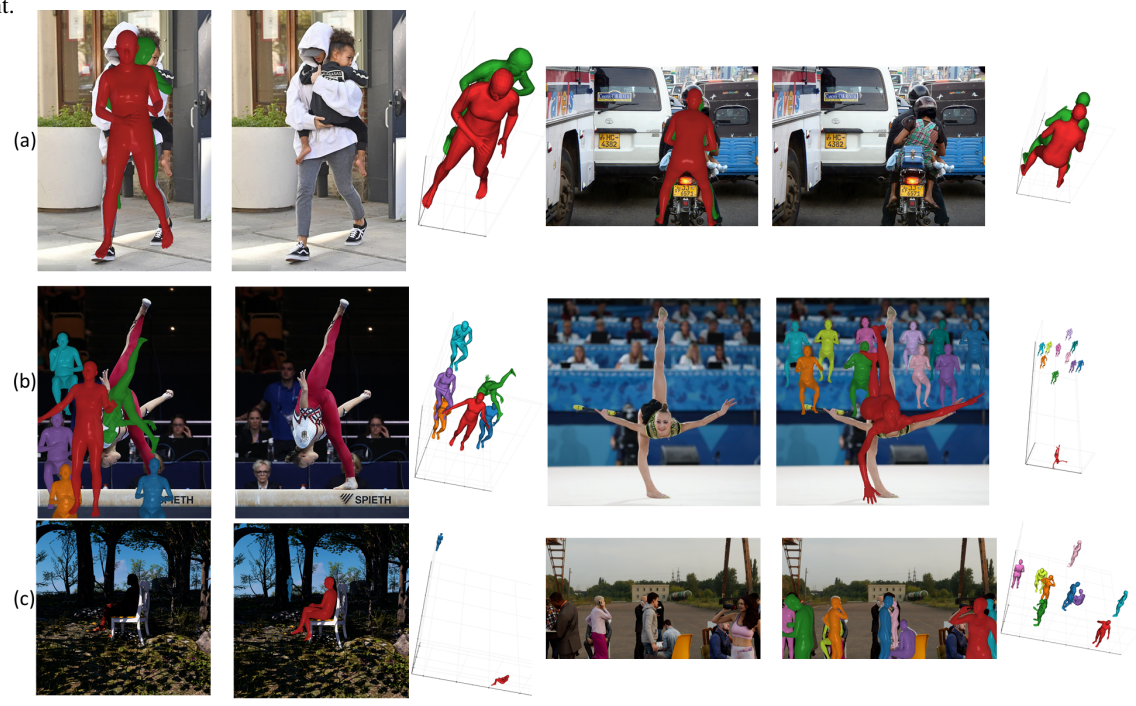

Figure 6. Example of failure cases.

Figure 7. Additional qualitative results of PLIKS from COCO [12], MPII [1], 3DPW [20], 3DOH [21] and MuPoTs-3D [14]. Set of challenging input images, overlay and, 3D view.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[2] Nikolas Hesse, Sergi Pujades, Javier Romero, Michael J. Black, Christoph Bodensteiner, Michael Arens, Ulrich G. Hofmann, Uta Tacke, Mijna Hadders-Algra, Raphael Weinberger, Wolfgang Müller-Felber, and A. Sebastian Schroeder. Learning an infant body model from RGB-D data for accurate full body motion analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2018.

[3] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2014.

[4] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Kalen Michael, Jiacong Fang, Imyhxy, , Lorna, Colin Wong, (Zeng Yifu), Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, Tkianai, YxNONG, Piotr Skalski, Adam Hogan, Max Strobel, Mrinal Jain, Lorenzo Mammana, and Xylieong. ultralytics/yolov5: v6.2 - yolov5 classification models, apple m1, reproducibility, clearml and deci.ai integrations, 2022.

[5] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In *3DV*, 2020.

[6] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131. IEEE Computer Society, 2018.

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[8] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proceedings International Conference on Computer Vision (ICCV)*, pages 11127–11137. IEEE, Oct. 2021.

[9] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11035–11045, Oct. 2021.

[10] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

[11] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. cite arxiv:1405.0312Comment: 1) updated annotation pipeline description and figures; 2) added new section describing datasets splits; 3) updated author list.

[13] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, pages 506–516. IEEE Computer Society, 2017.

[14] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018.

[15] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. *ArXiv*, abs/2008.03713, 2020.

[16] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS 2017 Workshop on Autodiff*, 2017.

[17] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. Agora: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478, June 2021.

[18] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[19] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, 2022.

[20] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, volume Lecture Notes in Computer Science, vol 11214, pages 614–631. Springer, Cham, Sept. 2018.

[21] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.