

IMPROVING 3D POSE ESTIMATION FOR SIGN LANGUAGE

Maksym Ivashechkin, Oscar Mendez, Richard Bowden

University of Surrey
Centre for Vision, Speech and Signal Processing (CVSSP)
Stag Hill, University Campus, Guildford GU2 7XH, UK

ABSTRACT

This work addresses 3D human pose reconstruction in single images. We present a method that combines Forward Kinematics (FK) with neural networks to ensure a fast and valid prediction of 3D pose. Pose is represented as a hierarchical tree/graph with nodes corresponding to human joints that model their physical limits. Given a 2D detection of keypoints in the image, we lift the skeleton to 3D using neural networks to predict both the joint rotations and bone lengths. These predictions are then combined with skeletal constraints using an FK layer implemented as a network layer in PyTorch. The result is a fast and accurate approach to the estimation of 3D skeletal pose. Through quantitative and qualitative evaluation, we demonstrate the method is significantly more accurate than MediaPipe in terms of both per joint positional error and visual appearance. Furthermore, we demonstrate generalization over different datasets and sign languages. The implementation in PyTorch runs at between 100-200 milliseconds per image (including CNN detection) using CPU only.

Index Terms— 3D pose estimation, hand and body reconstruction.

1. INTRODUCTION

Human pose estimation is an active and challenging field of research and recent years have seen significant progress in deep learning approaches to the estimation of 2D human key-points in images [1]. However, the breadth of potential applications, associated variability in appearance and the complexity of human motion make this an extremely challenging task. Lifting 2D to 3D (or estimating human pose directly in 3D) has inherent ambiguities which must be overcome using subtle visual cues. This often involves reasoning and/or complex statistical priors or 3D meshes [2].

In this work, we demonstrate that domain-specific training of pose regression outperforms existing general solutions, and we can trade generality for accuracy by focusing on the application. Specifically, we focus on sign language. However, even in the more general case, our simple approach can outperform the *state-of-the-art*.

Human pose estimation is an appealing first stage to support sign language recognition as a human skeleton provides natural invariance to a person, clothing, and background. However, sign language has many of its own challenges that lead to common failures for generic pose estimation techniques. During sign, the hands move quickly resulting in motion blur which leads to keypoint detection failure for frame-based pose estimation techniques. Furthermore, sign involves extensive hand-to-hand and hand-to-face interactions. These types of complex hand-to-body interactions are a common point of failure for pose estimation techniques. Fig. 1 gives two examples of OpenPose [3] and Monocular Total Capture [4] where



Fig. 1: Figure showing OpenPose failure for hand-to-hand interaction and associated incorrectly reconstructed mesh for Monocular Total Capture.

such hand-to-hand and hand-to-face interactions lead to failures in estimation, *e.g.*, ambiguous hand shape, distortion, or position.

This paper proposes a pose reconstruction pipeline that given a single image uses neural networks to generate human joint orientations and bone-length predictions which are fused with a kinematics model.

2. BACKGROUND

There are various techniques available for 3D pose estimation from a single image, which can be broadly categorized into two groups. The first group involves the direct regression of 3D points from the image using either 2D pose key-points or features obtained from an image processing model, such as a convolutional neural network (CNN). The second group of methods utilizes forward kinematics (FK) in combination with the uplift from 2D.

One design choice is whether to predict joint orientation via the networks rather than position. In QuaterNet [7] the authors predict 3D rotation via a quaternion representation. Zhou *et al.* [8] show that rotations in 3D have a continuous representation in at least five-dimensional space. In [9], Levinson *et al.* demonstrate that a 9D rotation matrix representation provides the best performance in model training compared to other representations. In [7–9] FK is applied to generate new poses after model prediction of rotations given a 3D pose input.

Human 3D pose estimation from a single image via Inverse Kinematics (IK) with an unscented Kalman filter was presented in [10], which shows better results than numerical IK reconstruction and faster convergence. Li *et al.* [11] exploit neural networks to predict 3D pose, then refine it using IK by running a twist-and-swing decomposition [12] to estimate the rotations of body joints. A single image approach was presented in [13] that predicts features and 2D

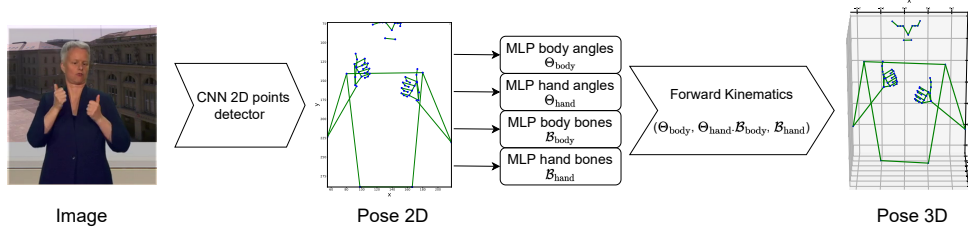


Fig. 2: The pipeline for pose reconstruction from a single image. The input image undergoes a 2D detection (*e.g.*, HRNet [5] or MediaPipe). Separate networks then generate rotations and bone lengths for the body and hands, and a custom FK layer in PyTorch combines this information to produce a 3D pose. Image source is DSGS EASIER dataset [6].

joints from an image, lifts them to 3D, and projects them back to the image to refine the estimate. Yang *et al.* [14] describe pose estimation from images combining 3D regression and pose prediction using Lie algebra with FK.

Pose reconstruction using a feed-forward network was demonstrated by Zhao *et al.* [15]. In the context of sign, pose estimation using quaternion prediction from 2D followed by FK was described by Krishna *et al.* [16]. In Elepose [17], Wandt *et al.* propose an unsupervised approach to 3D pose estimation (using only 2D poses input) by selecting the pose with the maximum likelihood over random 2D projections, where the likelihood is found via normalizing flows on 2D poses.

3. POSE RECONSTRUCTION

Our network structure is shown in Fig. 2. We use a CNN backbone to extract 2D key-points from an image, followed by four multi-layer perceptrons (MLPs) that generate joint Euler angles and bone lengths. Finally, a human pose is obtained via FK using the output angles and limb lengths.

A fully-connected linear network (MLP) is used to predict angles of human joints, and the FK propagates joint position and rotations from the root. The rotation is represented via Euler angles. We enumerate several reasons to justify this. Firstly, it is easier to represent a rotation matrix that has fewer than 3 degrees of freedom (DoF). This is needed because the human pose consists of many joints with only one or two DoF such as the elbows or fingers. Secondly, exploiting Euler angles helps to enforce additional limits and constraints for the rotation of human joints. For instance, the head can move around 90° degrees for roll, pitch, and yaw angles.

3.1. Angular network

The structure of the angular MLP includes several fully-connected layers (*e.g.*, 128-512 hidden units per layer) followed by a ReLU activation except for the last layer where a sigmoid function is used. The network’s inputs are 2D pose points flattened into a single vector. The sigmoid layer normalizes each output angle $\tilde{x} \in [0, 1]$ range, and the angular constraint is enforced by multiplying the angle’s limits, *i.e.*, $x = \tilde{x} \cdot (x_{\max} - x_{\min}) + x_{\min}$. This step not only guarantees the network outputs valid angles but also solves the problem of non-uniqueness and discontinuity (*e.g.*, 0 and 2π) of the Euler angle representation, since angles will always be lower than 180° corresponding to human limits. The root joint rotation has 3 DoF in the general case and remains unconstrained, *i.e.*, it has range $[-\pi; \pi]$ for roll, pitch, and yaw. This enables a human to rotate, perform a handstand, and lie on their side.

3.2. Limb lengths network

The bone length network has similar but smaller architecture than the angular network, because its role is less complex. We use an MLP with two layers in the experiments (see 4). The only constraint enforced for limb length is it being greater than zero, but sigmoid normalization can also be used to constrain bone lengths in the desired range.

3.3. Forward kinematics

The output of the angular network and the MLP for bone length are then combined in an FK layer to produce a 3D pose. The human pose is represented as a hierarchical tree structure where a root node is the root joint, *e.g.*, center of the hips. The position and rotation of the root joint indicate the origin of a pose and its orientation in the camera frame.

The FK is performed by traversing a tree \mathcal{T} with a breadth-first search algorithm [18]. The computation of each joint position \mathbf{p}_i and its rotation \mathbf{R}_i , $i > 0$ (except for the root’s index 0) is performed as $\mathbf{R}_i = \mathbf{R}_j \mathbf{R}'_i$ and $\mathbf{p}_i = \mathbf{R}_i \mathbf{o}_i + \mathbf{p}_j$ for each edge (i, j) of the graph \mathcal{T} , where j is a parent of i ; \mathbf{o}_i denotes the offset constructed from a bone length of a joint i , and \mathbf{R}'_i is the relative rotation of $0 \leq k \leq 3$ DoF computed from the predicted angles of the MLP. The rotation \mathbf{R}'_i with zero DoF equals an identity matrix. For instance, all points on a head (eyes, ears, etc) share only one rotation since they do not move independently. Hence, only one node has a full rotation describing the movement of the whole head while other points inherit this rotation as part of the FK. The orientation of the root joint equals to the relative rotation predicted by MLP since the root has no parent: $\mathbf{R}_0 = \mathbf{R}'_0$, and root location \mathbf{p}_0 (pose origin) defines translation for all other points.

3.4. Objective function

The angular MLP can be supervised by a single or combination of objective functions:

1. absolute difference w.r.t. the ground truth (GT) angles.
2. Euclidean distance between generated 3D and GT poses.
3. Euclidean distance between projected 3D and GT image points.

Denote $\mathbf{X}_W \in \mathbb{R}^{3 \times N}$ to be a pose of N points in columns in the world frame, and $\mathbf{P} = \mathbf{K}[\mathbf{R} \ \mathbf{t}]$ is a perspective projection matrix, where \mathbf{K} is the intrinsic camera matrix, and (\mathbf{R}, \mathbf{t}) are extrinsic parameters. The pose \mathbf{X}_W can be converted to pose $\mathbf{X}_C = \mathbf{R}\mathbf{X}_W + \mathbf{t}$ in the camera frame, and its projection onto image plane is $\mathbf{X}_I \sim \mathbf{K}\mathbf{X}_C$. Each of these objective functions has

advantages and disadvantages. The first loss, when applied to the predicted and the ground truth angles, provides faster training since FK is avoided. However, the MLP does not explicitly learn anything about the 3D point position, and errors on the root angles imply significantly higher errors on 3D points.

FK predicts poses in the camera frame, therefore, the loss in 3D is the Euclidean distance against the ground truth pose in the camera frame $\|\hat{\mathbf{X}}_C - \mathbf{X}_C^*\|$. In the experiments, two separate models for hands and body are used to avoid the fact that body pose errors are significantly higher than errors on hands, which can be problematic for training. This is useful as it allows the networks to be applied independently when only body or hands points are available.

A slightly worse performance on the validation set comes from the ground truth 2D points supervision when the reprojection error function is used (*i.e.*, $\|\hat{\mathbf{X}}_I - \mathbf{X}_I^*\|$), because a small change in estimated 3D pose results in a less accurate 2D projection with respect to the ground truth 2D points. Therefore, the network has to learn more about correct projections than a 3D pose, and it also implies significantly longer training.

The primary loss function used in the experiments is the Euclidean distance versus GT pose in 3D. Comparable performance is obtained using a combination of angular and 3D loss functions. The authors of [8, 9, 19] also use 3D distance for error computation, while in [7] the angular and pose losses are employed in different scenarios. The combination of 3D and reprojection loss with suitable weights as suggested by Kendal *et al.* [20] is the best option, because it enforces the model to learn both 3D points and their image correspondence, but it requires more exhaustive computations.

In summary, the human pose reconstruction method starts from the detection of 2D points from a single image using a CNN (MediaPipe in our experiments). Afterward, it exploits four sub-networks: two to predict joint angles and a further two for bone lengths of the body and hands. The proposed approach does not intend to predict the origin of a pose in 3D, because given only a single view this is a very challenging task, and it is not important for sign language.

The angular network and the MLP for bone lengths are trained separately to avoid the effect of one model on another. For the angular MLP training, the ground truth bone lengths and origin are used to generate a 3D pose. For the bone length MLP, if the GT angles are available, then the network also performs FK and calculates the error in 3D, otherwise, the loss is computed as the absolute difference to the ground truth bone values.

Since models are trained separately, a body is concatenated with hands afterward. It is important to add additional weights to the wrist points while computing the loss in 3D, otherwise, the error on the wrist joint implies a higher error on hand joints. The benefit of using a separate hand and body model is that only one MLP for both hands can be employed by simply flipping the input and output for the left hand.

4. EXPERIMENTS

For sign language, we train our model on the SMILE sign language dataset [21]. The ground truth for the SMILE sign language dataset was created by running IK with an Adam optimizer [22] using the PyTorch library. The optimization is performed over joint angles, bone lengths, and translations in 3D using calibrated multiview data provided by the dataset owners. To demonstrate the superior performance of the uplift process over generic approaches, we qualitatively compare against MediaPipe on three popular sign language datasets, SMILE dataset [21] which is Swiss German Sign Language (DSGS), the PHEONIX2014 dataset [23] which is German Sign Language

(DGS) and the BOBSL dataset [24] which is British Sign Language (BSL). The latter datasets were not seen/used during training and demonstrate excellent generalization across different sign languages. To provide a quantitative evaluation we first look at the accuracy on the SMILE dataset before retaining our model on the popular 3D pose estimation datasets PANOPTIC STUDIO [25], and HUMAN3.6M [26] demonstrating *state-of-the-art* performance.

We use MediaPipe 2D key-points of the upper torso (19 joints) as input to the lifting network since points below the waist are not relevant in the context of sign, and often not visible in an image. The body MLP generates in total 19 angles, *i.e.*, one for each elbow, 3 for each shoulder, 3 for head, 3 for the center of hips, 2 for the torso and 3 for the root, see Fig. 3 for visualization.

For PANOPTIC STUDIO and HUMAN3.6M datasets the input for the body MLP is the full MediaPipe body pose (33 points) and the target skeleton is the same as provided in datasets, *i.e.*, 19 points for the whole body for PANOPTIC and 17 joints for HUMAN3.6M. The number of angles predicted by the MLP is 29 for PANOPTIC and 31 for HUMAN3.6M. A mapping of angles to joints is similar to the sign dataset with new angles added for each leg.

The angular representation of a hand is adopted from [27], *i.e.*, in total one hand has 26 angles. The SMILE and PANOPTIC datasets have ground truth hands available, but HUMAN3.6M does not. Input for the hand MLP is again MediaPipe detection of the 21 joints of a single hand.

4.1. Quantative evaluation

The SMILE and PANOPTIC dataset were divided into training (80% of all video sequences), validation (10%), and test (10%) sets. For the HUMAN3.6M dataset five subjects are used for training and validation while the other two for testing, this split is commonly named as protocol #1 (see [28] or [13]).

MediaPipe 3D was obtained by running Holistic and Hands models separately, with parameters set to non-static image, 50% detection and tracking confidence (defaults values); and the highest model complexity available.

Table 1: Per-joint position errors on the SMILE dataset. \mathbf{R} , \mathbf{t} , s are rotation, translation, and scale applied to the predicted poses to align with the ground truth (GT). \mathcal{B} indicates GT bone lengths were used in FK, otherwise, limb lengths are estimated by the separate network. Columns show median error ('median'), MediaPipe confidence weighted average error ('w. mean'), and standard deviation ('std'). The row #fails shows the number of times MediaPipe failed to detect hands.

		SMILE	Per joint position error (cm)		
			median	w. mean	std.
MediaPipe	body	$\mathbf{R}, \mathbf{t}, s$	4.099	5.023	3.099
		\mathbf{t}, s	9.155	11.379	7.832
	hand	$\mathbf{R}, \mathbf{t}, s$	1.813	2.046	1.173
		\mathbf{t}, s	3.206	3.768	2.447
		#fails	443376 ($\approx 27.6\%$)		
Proposed	body	$\mathcal{B}, \mathbf{R}, \mathbf{t}$	0.928	1.427	1.584
		\mathcal{B}, \mathbf{t}	1.398	1.965	1.987
		\mathbf{R}, \mathbf{t}	1.126	1.649	1.647
		\mathbf{t}	1.556	2.134	2.040
	hand	$\mathcal{B}, \mathbf{R}, \mathbf{t}$	0.573	0.770	0.673
		\mathcal{B}, \mathbf{t}	0.833	1.143	1.044
		\mathbf{R}, \mathbf{t}	0.629	0.819	0.671
		\mathbf{t}	0.863	1.167	1.034

Table 2: Average, median, and standard deviation per-joint position errors (cm) on PANOPTIC and test subjects (S9 and S11) for HUMAN3.6M datasets. Individual rows show a specific alignment.

Dataset		Per joint position error (cm)			
		median	w. mean	std.	
PANOPTIC	body	$\mathcal{B}, \mathbf{R}, \mathbf{t}$	1.089	3.496	5.445
		\mathcal{B}, \mathbf{t}	1.421	4.378	5.353
		\mathbf{R}, \mathbf{t}	1.471	4.290	5.553
		\mathbf{t}	1.777	5.067	5.425
	hand	$\mathcal{B}, \mathbf{R}, \mathbf{t}$	0.558	1.012	1.402
		\mathcal{B}, \mathbf{t}	0.844	1.597	1.888
		\mathbf{R}, \mathbf{t}	0.593	1.052	1.399
		\mathbf{t}	0.874	1.635	1.889
H3.6M	body	$\mathcal{B}, \mathbf{R}, \mathbf{t}_{\text{root}}$	4.173	5.395	5.567
		$\mathcal{B}, \mathbf{t}_{\text{root}}$	4.914	6.429	6.839
		$\mathbf{R}, \mathbf{t}_{\text{root}}$	4.747	5.778	5.412
		\mathbf{t}_{root}	5.410	6.744	6.630

Results for the SMILE dataset are demonstrated in Table 1. Both our and MediaPipe results are aligned using Kabsch-Umeyama algorithm [29, 30], but our approach does not require scale alignment as bone lengths with approximately similar scale are predicted by the bone length MLP. Additionally, we report errors if the ground truth bone lengths are applied. From the Table 1, it can be seen the proposed network outperforms MediaPipe with significantly lower errors even if only a translational alignment is done.

The statistical results on the HUMAN3.6M dataset, including median and standard deviation of per-joint position error, are reported in Table 2 across all test sequences. Evaluation protocol #1 includes aligning a predicted pose to the ground truth root joint (pelvis point, see \mathbf{t}_{root} in the table). Accuracy on HUMAN3.6M is comparable to *state-of-the-art* which varies from 5.76 (best) to 16.21 (worst) [28].

The per-joint position errors for the PANOPTIC dataset are shown in Table 2. We report errors separately on body and hands, with or without using the ground truth bone lengths, and the Kabsch-Umeyama alignment. The errors on body pose are higher than for the sign language dataset, because actors in the PANOPTIC dataset demonstrate more complex motion. The results for hands are similar to the sign language dataset in terms of the median, but higher for the average. The reason for this is that the two hand models were trained separately, and the sign language dataset has better and more samples for training. Directly comparing these results against the *state-of-the-art* on PANOPTIC is challenging as papers such as [32] do not train on PANOPTIC only evaluate performance on this dataset. Despite this fact, if we look at the errors of Table 2 they are roughly half of those reported in [32]. A fairer comparison against [33] uses the same dataset and evaluation protocol where we reduce the error on body estimation by over 20mm.

4.2. Qualitative evaluation

A visual comparison of the proposed model and MediaPipe [34] is shown in Fig. 3 on various sign language datasets. Visually, poses predicted by the proposed method are better, especially, with respect

This work received funding from the SNSF Sinergia project 'SMILE II' (CRSII5 193686), the European Union's Horizon2020 research and innovation programme under grant agreement no. 101016982 'EASIER'. This work reflects only the authors view and the Commission is not responsible for any use that may be made of the information it contains.

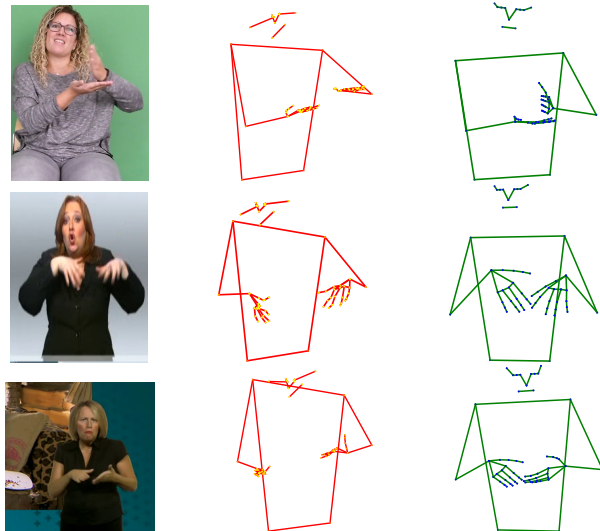


Fig. 3: Images show input signer (image left), MediaPipe output (red) in the middle, and our pose (green) in the right. From the top, we show SMILE dataset [21], middle RWTH-Phoenix Weather 2014 dataset [23] and bottom BBC-Oxford British Sign Language dataset [24, 31].

to the face and hands. MediaPipe hands' points are noisier without the preservation of limb length and orientation constraints. In many cases, MediaPipe hand detection completely failed for one or both hands.

5. CONCLUSIONS

This paper presented a 3D uplift approach that combines MLPs for 3D pose reconstruction from a set of 2D pose points in a single image. The primary contribution of this work is the combination of multiple prediction networks with a forward kinematic model that is able to generate valid and accurate 3D reconstructions. With a specific application to 3D pose estimation in sign language, the quantitative and qualitative evaluation shows that the method outperforms the commonly used MediaPipe 3D pose estimator in visual and accuracy tests. Furthermore, the model is capable of providing *state-of-the-art* performance on more general 3D pose estimation datasets.

6. REFERENCES

- [1] MMPose Contributors, "OpenMMLab pose estimation toolbox and benchmark," <https://github.com/open-mmlab/mmpose>, 2020. 1
- [2] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015. 1
- [3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "OpenPose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1
- [4] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh, "Monocular

- total capture: Posing face, body, and hands in the wild,” in *CVPR*, 2019. 1
- [5] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao, “Deep high-resolution representation learning for visual recognition,” 2019. 2
- [6] Maria Kopf, Marc Schulder, and Thomas Hanke, “Overview of Datasets for the Sign Languages of Europe,” July 2021. 2
- [7] Dario Pavlo, David Grangier, and Michael Auli, “Quaternet: A quaternion-based recurrent model for human motion,” 2018. 1, 3
- [8] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li, “On the continuity of rotation representations in neural networks,” *CVPR*, pp. 5738–5746, 2019. 1, 3
- [9] Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snively, Angjoo Kanazawa, Afshin Rostamizadeh, and Ameesh Makadia, “An analysis of SVD for deep rotation estimation,” in *Advances in Neural Information Processing Systems 34*, 2020. 1, 3
- [10] Yung-Ho Seo, Chil-Woo Lee, and Jong-Soo Choi, “Improved numerical inverse kinematics for human pose estimation,” *Optical Engineering - OPT ENG*, vol. 50, 03 2011. 1
- [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 *CVPR*, 2021. 1
- [12] Paolo Baerlocher and Ronan Boulic, *Parametrization and Range of Motion of the Ball-and-Socket Joint*, pp. 180–190, Springer US, Boston, MA, 2001. 1
- [13] Denis Tomè, Chris Russell, and Lourdes Agapito, “Lifting from the deep: Convolutional 3d pose estimation from a single image,” *CoRR*, vol. abs/1701.00295, 2017. 1, 3
- [14] Ji Yang, Youdong Ma, Xinxin Zuo, Sen Wang, Minglun Gong, and Li Cheng, “3d pose estimation and future motion prediction from 2d images,” *Pattern Recognition*, vol. 124, pp. 108439, 2022. 2
- [15] Ruiqi Zhao, Yan Wang, and Aleix Martinez, “A simple, fast and highly-accurate algorithm to recover 3d shape from 2d landmarks on a single image,” 2016. 2
- [16] Shyam Krishna, Vijay Vignesh, and Babu Dinesh, “Signpose: Sign language animation through 3d pose lifting,” in *ICCV*, October 2021, pp. 2640–2649. 2
- [17] Bastian Wandt, James J. Little, and Helge Rhodin, “Elepose: Unsupervised 3d human pose estimation by predicting camera elevation and learning normalizing flows on 2d poses,” *CoRR*, vol. abs/2112.07088, 2021. 2
- [18] Alan Bundy and Lincoln Wallen, *Breadth-First Search*, pp. 13–13, Springer Berlin Heidelberg, Berlin, Heidelberg, 1984. 2
- [19] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee, “Neural kinematic networks for unsupervised motion retargeting,” in *CVPR*, June 2018. 3
- [20] Alex Kendall and Roberto Cipolla, “Geometric loss functions for camera pose regression with deep learning,” *CVPR*, pp. 6555–6564, 2017. 3
- [21] Sarah Ebling, Necati Cihan Camgoz, Penny Boyes Braem, Katja Tissi, Sandra Sidler-Miserez, Stephanie Stoll, Simon Hadfield, Tobias Haug, Richard Bowden, Sandrine Tornay, et al., “SMILE Swiss German Sign Language Dataset,” in *Language Resources and Evaluation Conference*, 2018, number EPFL-CONF-233569. 3, 4
- [22] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014. 3
- [23] Oscar Koller, Jens Forster, and Hermann Ney, “Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers,” *Computer Vision and Image Understanding*, vol. 141, pp. 108–125, Dec. 2015. 3, 4
- [24] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman, “BOBSL: BBC-Oxford British Sign Language Dataset,” 2021. 3, 4
- [25] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh, “Panoptic studio: A massively multiview system for social interaction capture,” *TPAMI*, 2017. 3
- [26] 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.*, vol. 36, no. 7, pp. 1325–1339, 2014. 3
- [27] Ali Akbar Samadani, Dana Kulić, and Rob Gorbet, “Multi-constrained inverse kinematics for the human hand,” *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6780–6784, 2012. 3
- [28] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas, “Semantic graph convolutional networks for 3d human pose regression,” in *CVPR*, 2019, pp. 3425–3435. 3, 4
- [29] W. Kabsch, “A solution for the best rotation to relate two sets of vectors,” *Acta Crystallographica Section A*, vol. 32, no. 5, pp. 922–923, Sep 1976. 4
- [30] S. Umeyama, “Least-squares estimation of transformation parameters between two point patterns,” *PAMI*, vol. 13, no. 4, pp. 376–380, 1991. 4
- [31] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman, “BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues,” in *ECCV*, 2020. 4
- [32] 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*, June 2022. 4
- [33] Guillaume Rochette, Chris Russell, and Richard Bowden, “Weakly-supervised 3d pose estimation from a single image using multi-view consistency,” 2019. 4
- [34] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann, “Mediapipe: A framework for building perception pipelines,” *CoRR*, vol. abs/1906.08172, 2019. 4