Error bounds of projection models in weakly supervised 3D human pose estimation

Nikolas Klug, Moritz Einfalt, Stephan Brehm, Rainer Lienhart

Angaben zur Veröffentlichung / Publication details:

Klug, Nikolas, Moritz Einfalt, Stephan Brehm, and Rainer Lienhart. 2020. "Error bounds of projection models in weakly supervised 3D human pose estimation." In *2020 International Conference on 3D Vision (3DV), Fukuoka, Japan, 25-28 November 2020*, 898–907. Piscataway, NJ: IEEE. https://doi.org/10.1109/3DV50981.2020.00100.

Error Bounds of Projection Models in Weakly Supervised 3D Human Pose Estimation

Nikolas Klug[1*] Moritz Einfalt[2*] Stephan Brehm[2] Rainer Lienhart[2]
University of Augsburg
[1]klug.nikolas@gmail.com
[2]{firstname.lastname}@uni-a.de

Abstract

The current state-of-the-art in monocular 3D human pose estimation is heavily influenced by weakly supervised methods. These allow 2D labels to be used to learn effective 3D human pose recovery either directly from images or via 2D-to-3D pose uplifting. In this paper we present a detailed analysis of the most commonly used simplified projection models, which relate the estimated 3D pose representation to 2D labels: normalized perspective and weak perspective projections. Specifically, we derive theoretical lower bound errors for those projection models under the commonly used mean per-joint position error (MPJPE). Additionally, we show how the normalized perspective projection can be replaced to avoid this guaranteed minimal error. We evaluate the derived lower bounds on the most commonly used 3D human pose estimation benchmark datasets. Our results show that both projection models lead to an inherent minimal error between 19.3*mm and* 54.7*mm, even after alignment in position and scale. This is a considerable share when comparing with recent state-of-the-art results. Our paper thus establishes a theoretical baseline that shows the importance of suitable projection models in weakly supervised 3D human pose estimation.*

1. Introduction

The improvements in 2D human pose estimation (HPE) over the last years have been one of the most prominent successes of deep learning in computer vision tasks. Consequently, there is a still growing ambition to transition those advancements into the next logical step in human pose estimation: the monocular reconstruction of the 3D human pose, specifically from single images. Successful architectures and representations from the 2D task have been re-used to build end-to-end 3D pose reconstruction models with convincing results [16–18]. But the transfer into 3D space comes with additional difficulties that need to be addressed. First, labeled 3D human pose datasets are less available than large-scale 2D datasets. They are much more limited in scale as well as variability of appearance and poses. Second, the task of monocular 3D human pose reconstruction is less constrained and more difficult to optimize compared to its 2D counterpart [20]. Most prominently, this comes from the inherent ambiguities of object scale, depth and camera intrinsics. Different combinations can lead to the same 2D projection. This makes the reconstruction of the single, unique 3D pose impossible, unless strong priors about the object, scene and camera are given.

One approach to tackle both challenges can be seen in recent weakly supervised 3D HPE methods. Typically, they use two-stage pipelines, where initially the 2D human pose is inferred from the image using existing methods. Only this sparse 2D representation is fed into a separate deep neural network, which reconstructs the pose in 3D space. The 2D-

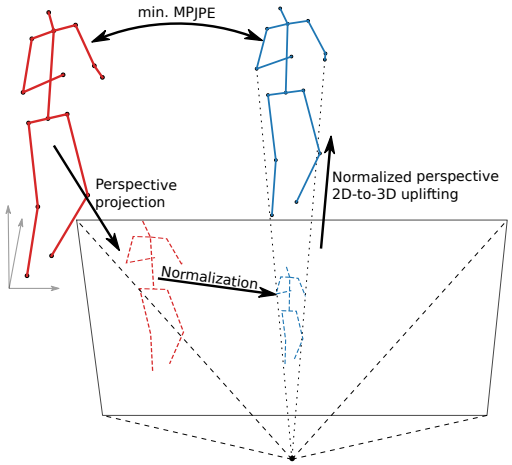

Figure 1: Lower bound error (MPJPE) for normalized perspective projection: We derive the best-case 2D-to-3D reprojection after 2D pose normalization and compare it to the original 3D human pose via MPJPE metric.

*Authors contributed equally

to-3D uplifting employs a generative approach, where the estimated 3D pose has to match a learned distribution of realistic human poses [3, 5, 25]. It avoids the need for direct supervision. One necessity is an explicit projection model to ensure a correct 3D-to-2D relationship.

Due to the ambiguities of monocular 3D reconstruction, current 2D-to-3D uplifting adopts simplified projection models. These add additional constraints that reduce the degrees of freedom of standard perspective projection with a pin-hole camera model. The two most common choices are the perspective projection after 2D normalization and the *weak perspective projection*. The former, which we refer to as the *normalized perspective projection*, simulates the perspective projection of a camera centered on the target human with a fixed projected scale. It enforces any input 2D pose to be artificially normalized in position and scale before 2D-to-3D uplifting [3]. The weak perspective projection model approximates the true perspective projection with a scaled orthogonal projection [10, 25]. Both variants do not match the original projection process in all but the most trivial cases. This mismatch negatively affects the fidelity of the 3D pose estimate.

In this paper, we provide a detailed analysis of those simplified projection models. Our goal is to derive a minimal error that these projection models are guaranteed to induce (Figure 1). For this, we develop a general formulation for the best-case 3D estimate that a 2D-to-3D uplifting network can find under those projections. By evaluating the best-case estimates on popular 3D HPE benchmark datasets, we provide a quantitative lower error bound for weakly supervised 2D-to-3D uplifting networks. It provides a baseline for the comparison of current fully and weakly supervised 3D human pose estimators. Additionally, we present a relaxed version of the normalized perspective projection model that avoids the necessity for position normalization. We integrate it into a baseline architecture and show how it can resolve the otherwise considerable error bound. Our contributions can be summarized as follows:

- We derive theoretical lower bound errors for weakly supervised 3D HPE under the constraints of normalized and weak perspective projection models.
- We evaluate the error bounds in synthetic experiments and on real-world 3D HPE benchmark datasets to identify the key factors that influence the minimal error.
- We propose a relaxed projection model for weakly supervised pipelines that can improve upon the original error bound by a large margin.

2. Related Work

Recent literature on deep learning based monocular 3D HPE is vast, with various paradigms, representations and architectures. We briefly cover recent work based on supervision type and the included projection models, if any, and relate it to our study. Note that our work focuses on 3D HPE in single images. There is highly active research on recovering temporal 3D pose trajectories from monocular videos (e.g. [7, 11, 12, 19, 27]), which we do not consider here.

Full supervision: Fully supervised approaches for translating images to 3D human poses are conceptually the most similar to state-of-the-art 2D HPE models. They usually build around fully convolutional architectures and spatially encoded regression targets [4, 16, 18, 23]. Limited by the need for sufficient 3D-annotated images, multiple proposals have been made to simplify the overall objective or relax the need for perfect 3D annotations. Kocabas *et al*. [13] use noisy 3D poses from multi-view epipolar geometry as the training input for their 3D HPE CNN. Rogez *et al*. [22] as well as Chen *et al*. [1] formulate 3D HPE as a selection task, where the best-fitting pose from a set of template poses is chosen. Apart from direct image translation, Martinez *et al*. [14] introduce a 2D-to-3D pose uplifting model with a fully supervised MLP. It effectively decouples the variability in human appearance from the much more constrained space of possible human poses. This paradigm has been prevalent in subsequently developed weakly supervised approaches. Overall, those fully supervised methods do not require manually projecting 3D pose estimates back into the 2D input domain, alleviating the requirement for an explicit projection model. In general, their performance is therefore not directly related to our established error bounds.

Weak and mixed supervision: Weakly supervised approaches for 3D HPE are usually characterized by 2D-to-3D pose uplifting in a generative framework. This is achieved with GAN-like networks which simultaneously learn the 3D pose reconstruction as well as its distribution. They enforce the 3D estimate to match the 2D input pose by integrating an explicit projection model. The 2D-to-3D uplifting process essentially re-projects the 2D input into 3D space according to the inverse projection model, while maintaining a realistic 3D estimate. Drover *et al*. [3] lift the normalized 2D input pose back into 3D space with respect to a depth estimate and a fixed perspective projection. Additionally, random projections of the resulting 3D estimate have to match a learned distribution of realistic 2D human poses. Since the normalized perspective projection is a structural component, it acts as a hard constraint and makes it directly applicable to our derived error bounds. Wandt *et al*. [25] propose a similar model, but instead directly learn a distribution of realistic 3D human poses in a kinematic representation. The consistency between the 3D estimate and the 2D input is ensured via weak perspective projection, but it only acts as a weak constraint that is part of the overall objective function to be optimized. Therefore, its performance is highly influenced but not necessarily bounded by our derived er-

ror bounds. Apart from purely weakly supervised methods, other end-to-end models with mixed supervision types have been established. Kanazawa *et al*. [10] try to estimate a parameterized 3D mesh of the human body in a generative fashion, while maintaining the image to 3D pose correspondence via direct 3D pose supervision and a weak perspective projection model. Similar to [25], the projection model also acts as a weak constraint. Finally, there are purely discriminative approaches with mixed 2D and 3D supervision. For example, Pavlakos *et al*. [17] map an estimated volumetric 3D heatmap representation to the scaled and cropped 2D input image via weak perspective projection. And even without the weak perspective projection, the unaccounted input image normalization alone would introduce the same issues as a normalized perspective projection. Therefore, such architectures are again directly related to our study.

Our work: We focus on weakly and mixed supervised methods that utilize simplified projection models. It is known that the objective function of monocular 3D human pose reconstruction is non-convex and difficult to optimize [20]. To the best of our knowledge, however, there is no existing study on lower error bounds in such approaches.

3. Error Bounds of Projection Models

Monocular 3D human pose estimation methods, either with 2D-to-3D pose uplifting or end-to-end image-based pipelines, often integrate a simplified projection model. The main intention is to reduce the degrees of freedom and to avoid ambiguities in the 3D estimation process [3, 25]. We show that this induces a systematical error into the 3D estimates and derive theoretical lower bounds for this error.

For the evaluation of 3D HPE on benchmark datasets, the most commonly used evaluation metric is the *Mean per-Joint Position Error* (MPJPE). It is calculated as the mean *Joint Position Error* (JPE; the Euclidean distances between ground truth and estimated joints) across all joints and poses. With unknown object size, depth and possibly camera intrinsics, estimating a correct absolute position and scale of the 3D human pose is inherently difficult. Common evaluation protocols therefore allow the alignment of 3D estimate and ground truth before calculating the MPJPE. The most relaxed variant is the *reconstruction error*, which allows the alignment of position, scale and rotation via Procrustes transformation. In this work we assume a more strict evaluation protocol that only allows alignment in position and scale. Specifically, we allow alignment of two designated root nodes and a per-subject scaling. Our approach for scaling is as follows:

1. Calculate the average mean limb length L_S of all poses of a subject (person) S.

2. For each subject S, scale the estimated 3D poses such that their mean limb lengths match L_S.

This procedure is commonly known as *Protocol 2* for evaluation on the Human3.6m dataset [8]. Identical or similar evaluation protocols are utilized in [14, 18, 24, 27, 28]. We specifically derive lower bounds for the MPJPE w.r.t. Protocol 2. Note that these minimal MPJPEs are optimistic best-case errors, such that the error bounds still hold for stricter protocols that only allow translation or no alignment at all. The error bounds will simply be less tight. Our derivations can be easily adjusted to better reflect stricter evaluation.

In the following sections, we work in the coordinate system of a standard perspective pin-hole camera with focal length f. The camera is located at the origin of the coordinate system and is oriented towards positive Z direction. We regard the 2D-to-3D human pose estimator G as a black box that, given a 2D input pose, predicts a reconstructed 3D pose such that its projection is identical to the input. In practice, this requires the estimator to at least predict an absolute depth value for each joint of the pose [3, 17]. The resulting 3D estimate is then given by inverting the projection model. In the following $P = [(X_1, Y_1, Z_1), \dots, (X_n, Y_n, Z_n)]$ denotes a 3D pose and $p = [(x_1, y_1), \dots, (x_n, y_n)]$ the corresponding 2D projection onto the image plane. In general, we assume f to be known.

3.1. Normalized Perspective Projection

In the normalized perspective projection model, the 2D pose is assumed to be normalized in position and scale. For 2D-to-3D pose uplifting, this necessitates translation and rescaling of the 2D input pose. We analyze the effects of both transformations on the final 3D estimate individually. Figure 1 gives an overview of our theoretical framework.

3.1.1 Effects of Translation

First, we consider the normalization of input 2D poses by translating them to the origin of the coordinate system. Note that this form of normalization is also part of many end-to-end 3D HPE pipelines that operate on cropped image patches. Our basic setup for analyzing the effects of pose translation is as follows: Let ground truth pose P be centered at the origin of the X-Y-plane at depth Z. Then the projection of joint P_i is given by $p_i = (x_i, y_i)$, with

$$x_i = f\frac{X_i}{Z_i}\ , \quad y_i = f\frac{Y_i}{Z_i}\ . \tag{1}$$

Now, we artificially shift P by the vector $(dx, dy, 0)$ and project it onto the image plane. After the resulting 2D pose is normalized by aligning it with the image plane's origin, it is re-projected into three dimensions by the 3D estimator G. The resulting 3D pose $\widetilde{P}$ is then compared to P via the MPJPE metric. For simplicity we assume that $dy = 0$, that is, the pose P is only shifted along the X-axis.

First, the artificial shift changes the X-component of the projected point p_i to

$$x_i^{\mathrm{S}} = f\frac{X_i + dx}{Z_i} = x_i + f\frac{dx}{Z_i}\,. \tag{2}$$

The projected pose is then normalized by shifting all projected points such that the root joint is located in the origin of the image plane. Let the shifted root joint P_r have coordinates $(dx, 0, Z)$. Since $dy = 0$, normalization shifts each each point by $-f\frac{dx}{Z}$ along the X-axis. The X-component $\widetilde{x}_i$ of the normalized point is given by

$$\widetilde{x}_i = x_i^{\mathrm{S}} - f\frac{dx}{Z} = x_i + f dx\left(\frac{1}{Z_i} - \frac{1}{Z}\right)\,, \tag{3}$$

while the Y-component remains the same as in Equation (1). After G estimates the depth $\widetilde{Z}_i$ for each joint, the 2D points are re-projected into three dimensions using standard perspective projection. The resulting coordinates are

$$\widetilde{X}_i = \frac{\widetilde{x}_i}{f}\cdot\widetilde{Z}_i = \widetilde{Z}_i\cdot\left(\frac{X_i}{Z_i} + dx\left(\frac{1}{Z_i} - \frac{1}{Z}\right)\right)\,, \tag{4}$$

$$\widetilde{Y}_i = \frac{y_i}{f}\cdot\widetilde{Z}_i = \frac{Y_i}{Z_i}\cdot\widetilde{Z}_i\,. \tag{5}$$

The JPE of the re-projected point is given by

$$\Delta d_i = \left\|\widetilde{P}_i - P_i\right\|_2\,. \tag{6}$$

With perfect estimation of $\widetilde{Z}_i = Z_i$, we have $\widetilde{Y}_i = Y_i$ and Equation (4) describes a linear relationship between the offset $|dx|$ and the JPE Δd_i. However, the JPE can actually become smaller for other $\widetilde{Z}_i \neq Z_i$. In order to obtain a correct lower bound for the JPE, we need to derive the $\widetilde{Z}_i$ which minimizes Δd_i. By substituting $a := \left(\frac{X_i}{Z_i} + dx\left(\frac{1}{Z_i} - \frac{1}{Z}\right)\right)$ and $b := \frac{Y_i}{Z_i}$, we have

$$\Delta d_i^2 = \left(\widetilde{Z}_i a - X_i\right)^2 + \left(\widetilde{Z}_i b - Y_i\right)^2 + \left(\widetilde{Z}_i - Z_i\right)^2 \tag{7}$$

This is a quadratic expression in $\widetilde{Z}_i$, whose global minimum is at

$$\widetilde{Z}_i^* = \frac{aX_i + bY_i + Z_i}{1 + a^2 + b^2} \tag{8}$$

Note that $\widetilde{Z}_i^*$ minimizes Δd_i^2 as well as the JPE Δd_i. The detailed derivation can be found in the supplementary material. Given, $\widetilde{Z}_i^*$, the resulting $\widetilde{X}_i^*$ and $\widetilde{Y}_i^*$ follow from Equation (4) and Equation (5). This yields a total minimal error of

$$\begin{aligned}\Delta d_i &= \sqrt{\left(\widetilde{X}_i^* - X_i\right)^2 + \left(\widetilde{Y}_i^* - Y_i\right)^2 + \left(\widetilde{Z}_i^* - Z_i\right)^2} \\ &= \left|dx\left(\frac{1}{Z_i} - \frac{1}{Z}\right)\right|\sqrt{\frac{Y_i^2 + Z_i^2}{1 + a^2 + b^2}}\,.\end{aligned} \tag{9}$$

The minimal MPJPE for the whole pose can then be calculated as the mean of all Δd_i. For a lower bound with $dx \neq 0$ *and* $dy \neq 0$, b has to be substituted with $\left(\frac{Y_i}{Z_i} + dy\left(\frac{1}{Z_i} - \frac{1}{Z}\right)\right)$ in the above equations.

The result shows that even with the system estimating the depth of each point optimally, the estimated 3D pose will not match the original 3D pose. The effects of 2D pose normalization by translation only vanish if $dx = 0$ or $Z_i = Z$ for all i. In the latter case all joints of the pose are located in a plane parallel to the image plane.

During evaluation with Protocol 2, the estimated 3D pose is aligned to the ground truth by shifting and scaling. For the root joint r we have $Z_r = Z$ in Equation (9), so $\Delta d_r = 0$. This means $(\widetilde{X}_r, \widetilde{Y}_r, \widetilde{Z}_r) = (X_r, Y_r, Z_r)$, that is, the root joints of both poses are already aligned. Additionally, Protocol 2 calculates a per-pose scaling factor $\alpha_{\widetilde{P}}$ such that the average limb length in $\alpha_{\widetilde{P}}\cdot\widetilde{P}$ is equal to the average limb length of the respective subject. Thus, for experimental evaluation, the pose $\widetilde{P}$ has to be re-scaled by $\alpha_{\widetilde{P}}$ in all dimensions before the MPJPE is calculated.

3.1.2 Effects of Scaling

It is common for many 3D HPE architectures to normalize the 2D input to match a fixed, but otherwise arbitrary scale. For 2D-to-3D uplifting, the 2D input pose is scaled to match *e.g.* a specific standard deviation or torso size [3, 25]. In image-based pipelines, this corresponds to resizing the input image to a specific scale. For a fixed 3D human pose, the scale of the 2D projection depends on the focal length f and the absolute depth Z. Since we assume f to be known, we analyze the effect of Z on the best-case 3D estimate.

Again consider points $P_i = (X_i, Y_i, Z_i) \in P$. Assume that P is shifted by dz along the Z-axis. The coordinates of the projected points on the image plane are

$$x_i = f\frac{X_i}{Z_i + dz}\,,\quad y_i = f\frac{Y_i}{Z_i + dz}\,. \tag{10}$$

Due to the Z-shift, the projected points now need to be scaled by ρ such that they match the fixed target scale. After G estimates the depths $\widetilde{Z}_i$, each point is re-projected into three dimensional space:

$$\widetilde{X}_i = \frac{\rho\cdot x_i}{f}\cdot\widetilde{Z}_i = \rho\cdot\frac{X_i}{Z_i + dz}\cdot\widetilde{Z}_i\,, \tag{11}$$

$$\widetilde{Y}_i = \frac{\rho\cdot y_i}{f}\cdot\widetilde{Z}_i = \rho\cdot\frac{Y_i}{Z_i + dz}\cdot\widetilde{Z}_i\,. \tag{12}$$

To obtain a best-case 3D estimate, we minimize the JPE with respect to $\widetilde{Z}_i$. Following Equation (6), the squared JPE for joint P_i is given by

$$\Delta d_i^2 = \left(\widetilde{Z}_i a - X_i\right)^2 + \left(\widetilde{Z}_i b - Y_i\right)^2 + \left(\widetilde{Z}_i - Z_i\right)^2\,, \tag{13}$$

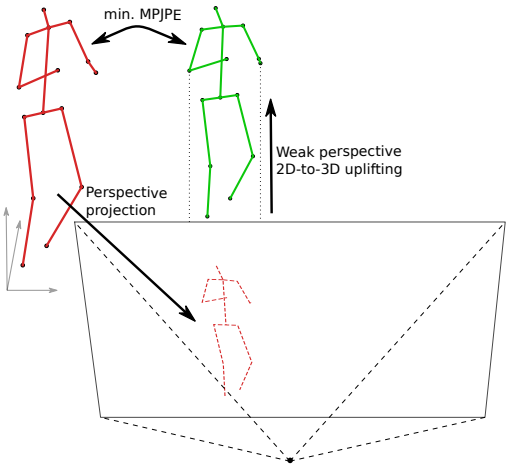

Figure 2: Lower bound MPJPE for weak perspective projection: We derive the best-case 2D-to-3D re-projection under this projection model and compare it to the 3D ground truth.

with $a \coloneqq \rho \cdot \frac{X_i}{Z_i+dz}$ and $b \coloneqq \rho \cdot \frac{Y_i}{Z_i+dz}$. Minimizing Δd_i^2 w.r.t. $\widetilde{Z}_i$ yields the same expression as in Equation (7), with

$$\widetilde{Z}_i^* = \frac{aX_i + bY_i + Z_i}{1 + a^2 + b^2} \,. \tag{14}$$

The detailed derivation is again part of the supplementary material. With $\widetilde{X}_i^*$ and $\widetilde{Y}_i^*$ from Equations (11) and (12), the minimal JPE is given by

$$\Delta d_i = \left|1 - \rho \cdot \frac{Z_i}{Z_i + dz}\right| \sqrt{\frac{X_i^2 + Y_i^2}{1 + a^2 + b^2}} \,. \tag{15}$$

The MPJPE Δd of the whole pose P can be calculated as the mean of all Δd_i. Similar to Section 3.1.1, during experimental evaluation, scaling and shifting according to Protocol 2 is applied.

The reason for the existence of an additional MPJPE is the scale factor ρ. As ρ is calculated for the whole pose rather than for each joint separately, $\rho \cdot \frac{Z_i}{Z_i+dz}$ is usually *not* equal to 1. This is only the case if $\rho = \frac{Z_i+dz}{Z_i}$ for all i, which means $Z_i = Z_j$ for all $1 \leq i, j \leq n$. In this case, all joints are located on the same X-Y-plane and the additional MPJPE vanishes.

3.2. Weak Perspective Projection

Instead of retaining a true perspective projection with normalized 2D inputs, many weakly and mixed supervised methods adopt the simpler weak perspective projection model. We can again derive an optimal 2D-to-3D human pose estimate under this projection model and, by comparing it to the 3D ground truth, establish a lower error bound. Figure 2 visualizes the conceptual setup.

Assume we have an arbitrary ground truth 3D pose P with joints $P_i = (X_i, Y_i, Z_i)$ and its 2D perspective projection p with joints $p_i = (x_i, y_i)$. The weak perspective projection model is structurally invariant to shifts perpendicular to the image plane. We therefore simplify our derivations by shifting p into the origin, resulting in p'. With p' as its input, the 3D estimator G can now predict the points in 3D space given as $\widetilde{P}_i = (\widetilde{X}_i, \widetilde{Y}_i, \widetilde{Z}_i)$. In order to reflect the X-Y shift in the input p', the goal is to reconstruct a equally X-Y centered version of P, denoted as P'. If the estimate matches P', it will equally match the original 3D pose P during Protocol 2 alignment. The 3D estimate has to adhere to a weak perspective projection, with

$$x'_i = f\frac{\widetilde{X}_i}{Z_{\text{avg}}} \,, \quad y'_i = f\frac{\widetilde{Y}_i}{Z_{\text{avg}}} \,. \tag{16}$$

Even though Z_{avg} is conceptually the average depth over all keypoints, it can be subsumed with f into an independent scaling factor $s = \frac{Z_{\text{avg}}}{f}$. The 3D estimator has to predict s alongside the individual depths $\widetilde{Z}_i$ for all keypoints [10,25]. With $\widetilde{Z}_i$ independent from s, the weak perspective model allows a best case depth estimate with $\widetilde{Z}_i = Z'_i$. We can now derive the optimal scaling s that minimizes the overall difference between $\widetilde{P}$ and P'. For a closed-form solution, we minimize the mean squared per-joint error Δd_2 instead of the non-squared MPJPE Δd:

$$\Delta d_2(s) = \frac{1}{n}\sum_i \left\| \begin{pmatrix} X'_i - sx'_i \\ Y'_i - sy'_i \\ Z'_i - \widetilde{Z}_i \end{pmatrix} \right\|_2^2 \,. \tag{17}$$

Under optimal depth estimation, Δd_2 is reduced to

$$\Delta d_2(s) = \frac{1}{n}\sum_i \left(X'_i - sx'_i\right)^2 + \left(Y'_i - sy'_i\right)^2 \,. \tag{18}$$

This quadratic expression is minimal at

$$s^* = \arg\min_s \Delta d_2(s) = \frac{\sum_i X'_i x'_i + Y'_i y'_i}{\sum_i x'^2_i + y'^2_i} \,. \tag{19}$$

The detailed derivation can again be found in the supplementary material. Note that in general, $s^* \neq \arg\min_s \Delta d(s)$, *i.e.* the obtained scale does not necessarily minimize the non-squared MPJPE. However, we have empirically found that using s^* only leads to an increase of $\Delta d(s)$ by less than 0.2mm compared to the actual optimal scaling on our evaluation datasets.

3.3. Relaxed Normalized Perspective Projection

In order to experimentally validate the derived lower bounds and to show how to circumvent them, we evaluate and modify a baseline 3D HPE architecture. For this we

consider the baseline 2D-to-3D HPE system by Drover *et al*. [3]. It is based on a weakly supervised GAN architecture [6] in which the generator receives 2D poses normalized in position and scale and estimates the depth of each joint. The 2D pose is then re-projected into 3D following standard perspective projection.

We re-implement this architecture with minor modifications: We remove batch normalization, as it proved detrimental for training stability and convergence and apply adaptive gradient clipping [2]. Apart from that, we use the same hyper-parameters.

Our experiments show that the effects of scaling might be negligible in most applications. Meanwhile, normalization by translation introduces a major increase in MPJPE. Thus, we focus on how to avoid the adverse effects of translation. In Equation (4), if the 2D points would be re-projected correctly (*i.e*. shifted back before re-projection), the lower bound in Equation (6) would no longer hold true. For this, assume the generator estimated all depths $\widetilde{Z}_i$ perfectly, that is, $\widetilde{Z}_i = Z_i$. If the 2D pose is shifted back and then re-projected into 3D, we have $\widetilde{X}_i = X_i$ and $\widetilde{Y}_i = Y_i$. In this case, the re-projected pose is now equal to the ground truth pose.

We adapt the baseline system by replacing the full normalization of the input 2D pose with two relaxed variants: In the first variant, the 2D input is still normalized in position and scale, but the re-projection into 3D space is performed on the non-shifted 2D input (thus invalidating the lower bound). For the second variant, the normalization by translation is removed entirely. This allows the 3D estimator to condition its per-joint depth estimate on the actual, non-shifted 2D joint locations. In both cases, the system is explicitly trained on uniformly shifted 2D input poses.

4. Experimental Results

We first show how the derived lower bounds for the normalized perspective projection change with varying offsets of artificially shifted poses. We validate them by comparing them to the baseline 2D-to-3D pose estimation system in [3]. Additionally, we evaluate the proposed relaxed normalization when integrated into the same 3D pose estimator. Finally, in order to quantify the theoretical error bounds' impact, we evaluate them for the normalized and weak perspective projection on the test sets of three commonly used 3D HPE benchmark datasets, namely Human3.6m [8], MPI-INF-3DHP [15] and CMU Panoptic [9]. Note that in all experiments, we assume perfect 2D poses as the input to 2D-to-3D uplifting. All results are evaluated according to Protocol 2, unless stated otherwise.

Human3.6m [8] consists of 3D human poses captured by a MoCap system and four RGB cameras in an indoor lab setting. We use the common split with subjects S9 and S11 for evaluation and the remaining subjects for training and employ the same 14 joint model as [3] with the addition of the pelvis. Most of our experiments are conducted on this dataset, as it is the largest and most commonly used benchmark for single-person 3D HPE. Similarly, *MPI-INF-3DHP* [15] contains 3D human poses from different activities, but with more viewpoint and pose variety. We use the pre-defined split and evaluate on all test subjects. Finally, *CMU Panoptic* [9] is the currently largest multi-person 3D HPE dataset. The poses are captured by 30 HD cameras arranged in a sphere-like structure. We treat the poses of all subjects individually and adopt the evaluation protocol from [4, 26], excluding the no longer maintained "Mafia" sequence.

4.1. Translation

First, we analyze the effect of translation on the normalized perspective lower bound error and compare it to the observed performance of the corresponding baseline system in [3], trained on Human3.6m. We use all 3D test set poses in the Human3.6m dataset, center them in the X-Y plane and place them on the Z axis such that the resulting 2D projections already satisfy the target scale of the baseline system. This results in an average camera distance of 5m to the poses. We then artificially shift the 3D poses along the X axis by offset dx. After projecting them onto the image plane, we calculate the best-case 3D estimates following Equation (8) and evaluate them according to Protocol 2. Additionally, we feed the synthetic 2D poses into the baseline 3D pose estimation system. The results with respect to dx are depicted in Figure 3.

Starting with its minimum at $dx = 0$m, the theoretical lower bound error quickly increases nearly linearly in the depicted interval, with a minimal MPJPE of already 62.9mm at only $dx = 2$m offset. At an average depth of 5m, this reflects a still very common camera viewpoint, where the human pose is observed at an angle of $22°$. The results of the baseline 2D-to-3D pose estimator show similar characteristics. Starting at an MPJPE of 55mm at $dx = 0$m, it increases similarly with a doubled MPJPE at only 2m offset. For larger offsets, the error increases even faster than the lower bound error. Overall, the results show that normalization by translation leads to a quickly increasing lower bound error, which can be equally observed for the baseline 3D estimator.

4.2. Scaling

We use the same experimental setup to evaluate the theoretical lower bound error that is introduced by scale normalization. This time, however, we artificially shift the 3D poses by varying offsets dz along the optical axis. This results in 2D projections that need to be normalized in scale to match the target scale of the baseline system. We evaluate the theoretical error as well as the change in MPJPE when

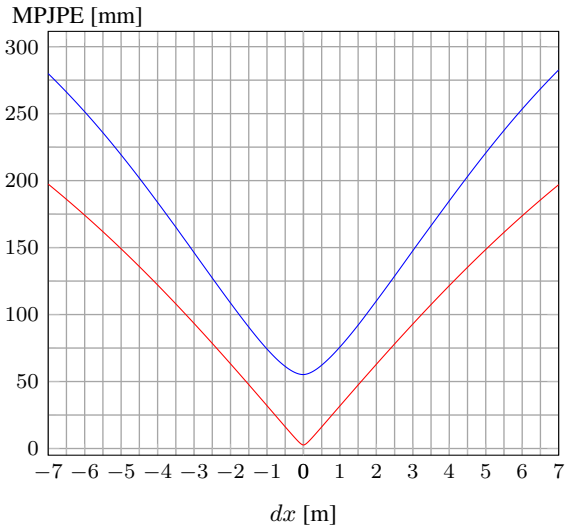

Figure 3: Theoretical lower bound (red) and experimental (blue) MPJPE with [3] on Human3.6m poses at different offsets dx. The theoretical error evaluates the best-case 3D estimates under normalized perspective projection.

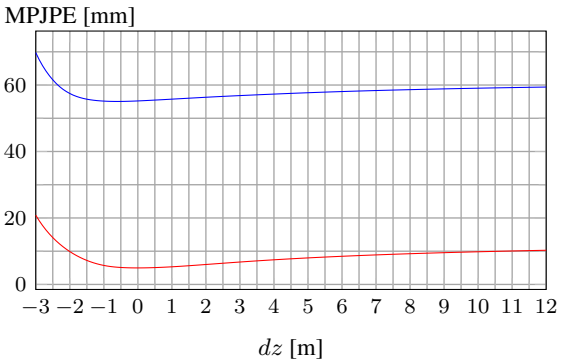

Figure 4: Theoretical lower bound (red) and experimental (blue) MPJPE with [3] on Human3.6m poses at different depth offsets dz. The depth at $dz = 0$ is approximately 5m.

applying the baseline system. The results are depicted in Figure 4.

The behavior of the theoretical error and the baseline system again show similar characteristics. Compared to the effect of shifting along the X axis, the MPJPE does not grow significantly for changes in dz. Only when poses are moving closer to the camera we observe a notable MPJPE increase, with 20mm at -3m offset. This results from the fact that with smaller camera distances, projection ray angles change more severely. The reason for the 5mm theoretical MPJPE at $dz = 0$ stems from Protocol 2 evaluation. As there is intra-subject variance in average limb lengths in Human3.6m, this minor error is introduced when scaling the estimated poses.

In general, the results show that normalization by scaling does introduce a systematic error. Apart from the case of poses being very close to the camera, the increase in MPJPE might be negligible in most applications.

4.3. Relaxed Normalization

Given the notable influence of normalization by translation, we evaluate our proposed relaxed normalization with the same experimental setup as in Section 4.1. Again, the baseline as well as the proposed variants are trained on Human3.6m.

The results for varying offsets dx are depicted in Figure 5. With the re-projection into 3D space based on the original 2D input, the MPJPE increases much slower with respect to dx and clearly surpasses the lower error bound in Figure 3. Compared to the baseline, we observe a negative effect only for very small offsets dx. Avoiding normalization by translation entirely leads to another reduction of the MPJPE by 10mm across all offsets. At $dx = 0$ the performance now equals the baseline system. This confirms the conjecture that the depth estimates need to be conditioned on the offsets dx to avoid ambiguities in the 3D estimates. At a large offset of $dx = 7$m, the relaxed normalization leads to an increase in MPJPE of only 6.6mm, compared to the 227.4mm for the baseline. Despite no normalization by translation, we do not observe any divergence issues.

We also evaluate on the actual Human3.6m test set, *i.e.* the 3D and 2D poses as observed by the camera setup. Table 1 (top) shows the resulting MPJPE under Protocol 2 for the baseline system and our proposed modification with relaxed normalization. At 50.9mm, our approach improves upon the baseline by notable 15.4mm. Despite the limited variability in the Human3.6m poses and viewpoints, the baseline is still heavily influenced by translation normalization. We additionally evaluate the less strict reconstruction error, where we observe identical results for the baseline as well as our relaxed normalization variant. The systematic error of translation normalization for 2D-to-3D uplifting is compensated by the additional orientation alignment during evaluation. For applications where pose orientation relative to the camera is irrelevant, the effects of translation are thus far less severe.

Overall, our simple modifications to an otherwise unchanged 2D-to-3D uplifting model can avoid the adverse effects of normalization almost entirely. Our proposal can be integrated into every weakly supervised system following the same normalized perspective projection.

4.4. Lower Error Bounds on 3D HPE Datasets

Finally, Table 1 (bottom) shows the evaluated lower bound errors for the normalized and the weak perspective projection model on the test set poses of the most common

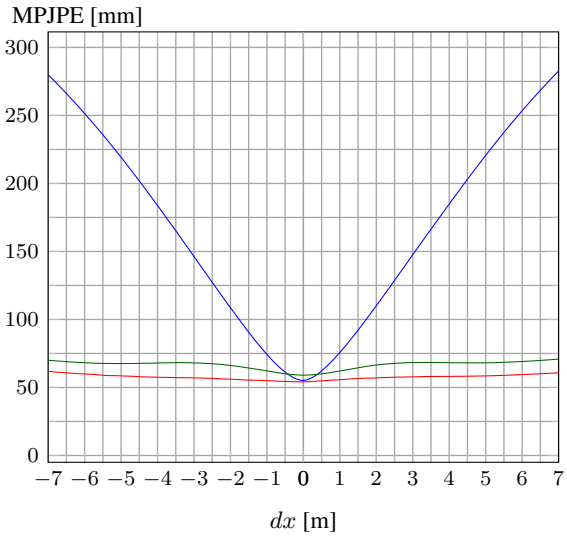

Figure 5: Comparison of the baseline [3] (blue) and our modified version with relaxed normalization for Human3.6m poses at different offsets dx: correct re-projection (green); without normalization by translation (red).

	Protocol 2	Reconstruction error
[3]	66.3	39.0
Ours	50.9	39.0

	Human 3.6m	MPI-INF-3DHP	Panoptic
State-of-the-art	49.6 [23]*	101.5 [21]	60.7 [4]*
Norm. persp.	19.3	52.0	54.7
Weak persp.	20.6	47.8	47.2

Table 1: Top: Results of 2D-to-3D uplifting with relaxed normalization on Human3.6m. Bottom: Comparison of state-of-the-art results and our theoretical lower bounds evaluated with Protocol 2 on common 3D HPE benchmark datasets. All results are given in MPJPE (mm). * use stricter evaluation without scale alignment.

3D HPE benchmarks. For the normalized perspective variant, we only assume normalization by translation due to its considerable negative effect. Additionally, Figure 6 depicts qualitative examples of the best-case pose estimates under those projection models. We provide additional examples in the supplementary material.

Across all datasets, both projection models lead to a similar lower bound MPJPE. For Human3.6m, the minimal error is relatively small, at around 20mm. For MPI-INF-3DHP and CMU Panoptic, however, the minimal error is at approximately 50mm already. Compared to Human3.6m, those datasets show a much wider variety in camera viewpoints, with poses being off-center, closer to the camera and with more absolute depth variation. This seems to affect the best-case 3D estimates with normalized and weak perspective projection to the same degree. Since the latter datasets better reflect in-the-wild human 3D poses, the effect of simplified projection models on 3D HPE applications is considerable. For reference, we report current state-of-the-art results on the respective datasets. Note that these results are obtained with fully supervised discriminative methods that do not employ simplified projection models. They are not limited by our derived lower bounds. Still, when comparing the results, the derived minimal errors are already at 50% or more of the state-of-the-art MPJPE results. Note that the minimal errors assume otherwise maximally optimistic 3D estimates with perfect 2D input poses. Therefore, any approach that employs 2D-to-3D uplifting with either normalized or weak perspective projection will hardly be able to surpass the current state-of-the-art.

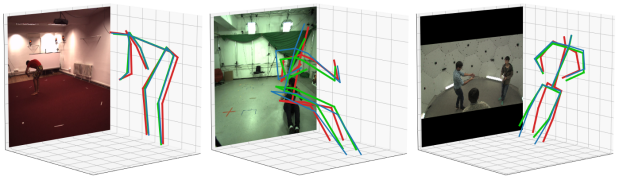

Figure 6: Exemplary ground truth poses (red) from Human3.6m, MPI-INF-3DHP and CMU Panoptic (left to right). The best-case 3D estimates under normalized and weak perspective projection are depicted in blue and green.

5. Conclusion

We presented a theoretical analysis of simplified projection models that are used in current weakly supervised 2D-to-3D HPE uplifting methods. We derived lower bound errors under a common evaluation protocol and metric by providing closed-form solutions for the best-case 3D pose estimates. Our analysis showed that 2D pose normalization by translation as well as the weak perspective projection model introduce a considerable error, even under overly-optimistic conditions. We showed how a relaxed normalization can be directly integrated into a baseline architecture to overcome the otherwise guaranteed minimal error. The evaluation on real-world 3D HPE datasets provides a baseline for future comparison between weakly and fully supervised methods. The comparison to state-of-the-art results showed that it is crucial for any weakly supervised architecture to avoid simplified projection models in order to achieve competitive results on current benchmarks. Our proposed relaxed normalization is one possibility to consider.

References

[1] 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 (CVPR)*, July 2017. 2

[2] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio. End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results. In *NIPS 2014 Workshop on Deep Learning*, Dec. 2014. 6

[3] D. Drover, R. MV, 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) Workshops*, September 2018. 2, 3, 4, 6, 7, 8

[4] M. Fabbri, F. Lanzi, S. Calderara, S. Alletto, and R. Cucchiara. Compressed Volumetric Heatmaps for Multi-Person 3D Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7204–7213, 2020. 2, 6, 8

[5] H.-Y. Fish 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. June 2014. 6

[7] A. Grinciunaite, A. Gudi, E. Tasli, and M. den Uyl. Human Pose Estimation in Space and Time using 3D CNN. In *Computer Vision – ECCV 2016 Workshops*, Oct. 2016. 2

[8] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, Jul 2014. 3, 6

[9] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015. 6

[10] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-End Recovery of Human Shape and Pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3, 5

[11] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019. 2

[12] M. Kocabas, N. Athanasiou, and M. J. Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 2

[13] M. Kocabas, S. Karagoz, and E. Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1086, 2019. 2

[14] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. *2017 IEEE International Conference on Computer Vision (ICCV)*, Aug. 2017. 2, 3

[15] 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 International Conference on 3D Vision (3DV)*, Oct. 2017. 6

[16] A. Nibali, Z. He, S. Morgan, and L. Prendergast. 3d human pose estimation with 2d marginal heatmaps. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1477–1485. IEEE, 2019. 1, 2

[17] G. Pavlakos, X. Zhou, and K. Daniilidis. Ordinal Depth Supervision for 3D Human Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 3

[18] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1263–1272, 2016. 1, 2, 3

[19] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. 3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[20] A.-I. Popa, M. Zanfir, and C. Sminchisescu. Deep Multitask Architecture for Integrated 2D and 3D Human Sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 3

[21] 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 (CVPR)*, June 2018. 8

[22] G. Rogez and C. Schmid. MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3108–3116. Curran Associates, Inc., 2016. 2

[23] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei. Integral Human Pose Regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2, 8

[24] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua. Direct Prediction of 3D Body Poses from Motion Compensated Sequences. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 991–1000, 2015. 3

[25] B. Wandt and B. 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. 2, 3, 4, 5

[26] A. Zanfir, E. Marinoiu, and C. Sminchisescu. Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes - The Importance of Multiple Scene Constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 6

[27] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4966–4975, 2015. 2, 3

[28] X. Zhou, M. Zhu, G. Pavlakos, S. Leonardos, K. G. Derpanis, and K. Daniilidis. MonoCap: Monocular Human Motion Capture using a CNN Coupled with a Geometric Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:901–914, 2017. 3