MeTRAbS: Metric-Scale Truncation-Robust Heatmaps for Absolute 3D Human Pose Estimation

István Sárándi, Student Member, IEEE, Timm Linder, Member, IEEE, Kai Oliver Arras, Member, IEEE, and Bastian Leibe, Member, IEEE

Abstract—Heatmap representations have formed the basis of human pose estimation systems for many years, and their extension to 3D has been a fruitful line of recent research. This includes 2.5D volumetric heatmaps, whose X and Y axes correspond to image space and Z to metric depth around the subject. To obtain metric-scale predictions, 2.5D methods need a separate post-processing step to resolve scale ambiguity. Further, they cannot localize body joints outside the image boundaries, leading to incomplete estimates for truncated images. To address these limitations, we propose metric-scale truncation-robust (MeTRaB) volumetric heatmaps, whose dimensions are all defined in metric 3D space, instead of being aligned with image space. This reinterpretation of heatmap dimensions allows us to directly estimate complete, metric-scale poses without test-time knowledge of distance or relying on anthropometric heuristics, such as bone lengths. To further demonstrate the utility of our representation, we present a differentiable combination of our 3D metric-scale heatmaps with 2D image-space ones to estimate absolute 3D pose (our MeTRaB architecture). We find that supervision via absolute pose loss is crucial for accurate non-root-relative localization. Using a ResNet-50 backbone without further learned layers, we obtain state-of-the-art results on Human3.6M, MPI-INF-3DHP and MuPoTS-3D. Our code is publicly available.

István Sárándi and Bastian Leibe are with the Chair of Computer Vision, RWTH Aachen University, 52062 Aachen, Germany (e-mail: sarandi@vision.rwth-aachen.de; leibe@vision.rwth-aachen.de).
Timm Linder and Kai Oliver Arras are with Robert Bosch GmbH, Corporate Research, 71272 Renningen, Germany (e-mail: timm.linder@de.bosch.com; kaioliver.arras@de.bosch.com).
Digital Object Identifier 10.1109/TBIOM.2020.3037257
https://vision.rwth-aachen.de/metrabs

Fig. 1. By defining heatmaps in the 3D metric space around the person (top row) we directly estimate scale-correct poses. This is in contrast to prior work (bottom row) that defines the X and Y heatmap axes in image space and requires further post-processing to obtain a metric-scale skeleton. The three columns show that this heatmap representation is nearly invariant w.r.t. image zooming. A knee heatmap is shown along with the soft-argmax decoded skeleton.

Manuscript received July 10, 2020; revised September 30, 2020; accepted November 6, 2020. Date of publication November 11, 2020; date of current version February 12, 2021. This work was supported in part by the Bosch Research Foundation; in part by the European Research Council under Consolidator Grant Project DeeViSe ERC-CoG-2017-773161; in part by the EU through the H2020 Projects under Grant ILIAD H2020-ICT-2016b-732737 and Grant CROWDBOT H2020-ICT-2017-779942. This article was recommended for publication by Associate Editor A. A. Salah upon evaluation of the reviewers’ comments. (Corresponding author: István Sárándi.)

Heatmaps are a commonly used representation to depict 2D human poses. They have been shown to be well suited for human pose estimation and are widely used in computer vision research. In this work, we propose a differentiable combination of metric 3D heatmaps with 2D image-space ones to estimate absolute 3D pose. This method is nearly invariant to image zooming and outperforms previous approaches. The code is publicly available.

I. INTRODUCTION

Human pose estimation is a long-standing computer vision problem with wide applicability in human-robot interaction [1], virtual reality [2], medicine [3] and commerce [4], among others. Since the adoption of deep convolutional neural networks (CNN), and especially heatmap representations, we have witnessed rapid progress in pose estimation research [5], [6], [7]. A particularly challenging task is monocular 3D pose estimation [8], [9], [10], [11], [12], where a person’s anatomical landmarks are sought in 3D space, i.e., in millimeters, instead of pixels, given only a single image. Reconstructing 3D from images is one of the major goals of computer vision research, but several geometric ambiguities make this challenging. First, different 3D articulations may share the same 2D projection and second, there is an ambiguity between object size and distance, as small objects near the camera appear the same as large ones far away.

There is no clear consensus on the most effective way to represent and tackle these problems. Heatmap estimation is a promising direction, because it makes direct use of the convolutional structure of CNNs by turning the coordinate estimation problem into a binary classification problem of whether the joint is located at the given position or not. To estimate 3D pose, a successful line of works extends 2D joint heatmaps with a depth axis, resulting in a 2.5D volumetric representation [13], [14], [15], [16].

Finding heatmap maxima gives the estimated pixel coordinates and root-relative metric depths per joint (a 2.5D pose). While these estimates can be accurate, 2.5D representations do...
not address the challenging ambiguity between person size and distance. To bridge the gap between a 2.5D and a 3D pose, a separate scale recovery step is needed in post-processing. Explicit anthropometric heuristics have been proposed as scale cues, e.g., bone length priors [13] or a skeleton length prior [17], computed by averaging over the training poses. However, these simple heuristics have difficulties when the experimental subjects have diverse heights. A further limitation is that 2.5D formulations are constrained to the estimation of joints that lie within the image boundaries. This is problematic in practical applications, where the image crop may not include the whole person, e.g., due to occlusions or detector noise. While one could use an additional module to fill in missing joints, it is desirable to learn the complete skeleton estimation in a single unified stage.

Our goal in this article is to tackle scale and distance estimation of 3D poses in a truncation-robust, simple and efficient manner, while also keeping the structural advantages of fully-convolutional heatmap estimation, as opposed to numerical coordinate regression (i.e., encoding position by activation location instead of activation value).

To this end, we propose training a fully-convolutional network to output our novel metric-scale truncation-robust (MeTRo) heatmaps as illustrated in Fig. 1. All dimensions of these heatmaps are defined to have a fixed metric extent in meters. This is an unconventional task definition for fully-convolutional networks (FCN). FCNs are predominantly applied for pixel-wise prediction tasks, such as semantic segmentation, where the input and output are pixel-to-pixel aligned. In our proposed approach, the input pixel positions and the output metric positions only satisfy a looser form of spatial correspondence. Nevertheless, we show that somewhat surprisingly, such a mapping can still be learned effectively by a standard modern FCN backbone.

By skipping the 2.5D stage, the backbone FCN has to implicitly reason about out-of-image joints, discover scale cues and learn the geometric perspective back-projection in an end-to-end manner. Our MeTRo heatmaps can naturally encode body parts outside the image, since the prediction volume’s bounds do not correspond to the image bounds. As there is no need to design an explicit scale recovery step, the pipeline becomes simpler and requires neither the focal length nor the root joint distance to be known at test time for root-relative prediction.

Employing the differentiable soft-argmax [14], [16], [18], [19] layer, our method becomes end-to-end learned all the way from image to final 3D metric-scale prediction as shown in Fig. 2. Soft-argmax also allows rapid training with low-resolution heatmaps. Without any additional learned decoder module, we perform dense prediction with reduced strides at test time for higher quality results. We find that the details of the striding mechanism are crucial and propose a centered striding method that distributes the output neuron receptive fields evenly over the image.

This article presents an extension of our own previous work [20]. While in [20] we only considered single-person root-relative pose, here we show that MeTRo heatmaps are also effective for absolute (non-root-relative) 3D pose estimation. In multi-person scenes it is especially important to estimate absolute poses, in order to recover the spatial layout of the whole group. We combine 3D metric-scale root-relative heatmaps with 2D image-space heatmaps in a two-headed CNN architecture, and subsequently reconstruct the absolute 3D root position in a differentiable manner. While prior approaches have tackled the root reconstruction problem, to our knowledge we are the first to backpropagate gradients through this reconstruction, allowing us to end-to-end supervise the absolute pose task. We evaluate our network in a top-down fashion combined with an off-the-shelf person detector. We refer to this combined approach as MeTRAbs.

Recent approaches have achieved good generalization performance to in-the-wild images by using abundant and diverse images with 2D pose labels in the training procedure besides 3D data [10], [14], [16]. Applying such weak supervision is challenging in our representation, since the MeTRo heatmap would require supervision with metric ground truth instead of the image-space ground truth supplied with 2D datasets. We tackle this by proposing a scale and translation agnostic loss for 2D-annotated examples using an alignment layer. Note that in contrast to 2.5D heatmaps, this alignment is only used for loss computation during training, and still allows MeTRo to infer joints outside the image boundaries.

Experimentally, we achieve state-of-the-art results on the two largest single-person 3D pose benchmarks, Human3.6M and MPI-INF-3DHP, as well as the popular multi-person dataset MuPoTS-3D. To isolate the effect of the representation, we perform direct comparisons with 2.5D heatmap learning using bone-length-based scale recovery [13], under otherwise equal training conditions. We find that scale cues can indeed be learned implicitly in this fashion and MeTRo outperforms the baseline on most test sequences. Furthermore, our approach achieved first place in the 2020 ECCV 3D Poses in the Wild [21] Challenge.

In summary, we make the following contributions: 1) We propose a novel 3D heatmap representation for pose estimation, called MeTRo, whose dimensions are defined in a fixed metric scale, irrespective of the input image scale. We achieve state-of-the-art results on Human3.6M and MPI-INF-3DHP and demonstrate strong truncation-robustness. 2) We propose centered striding, an improvement to the usual CNN striding logic, enabling higher accuracy at a coarse (8 × 8) heatmap resolution. 3) For absolute pose estimation, we extend the MeTRo approach to MeTRAbs, by also estimating 2D image-space heatmaps from the same backbone and reconstructing the absolute pose. We achieve state-of-the-art results on the MuPoTS-3D and 3DPW multi-person benchmarks using MeTRAbs in a top-down paradigm. 4) To our knowledge, we are first to use a monocular geometry-based differentiable absolute pose reconstruction module to supervise the network with the final absolute ground truth fully end-to-end. We show that this is crucial for good distance estimation and extensively evaluate strong and weak perspective-based reconstruction variants.
II. RELATED WORK

3D human pose estimation has had a long research history starting with hand-crafted features and part-based models [22]. Similar to other computer vision problems, the transition to deep convolutional networks has led to a dramatic performance increase in this task. For a thorough overview, see the recent survey by Chen et al. [23].

A. Deep 3D Human Pose Estimation

Much of the inspiration in recent 3D pose estimator design has come from lessons learned in 2D pose research. DeepPose, the first neural 2D pose estimator [24] directly regressed 2D joint coordinates on the RGB input via convolutional and fully-connected layers. Later, top-performing 2D methods have transitioned to predicting body joint heatmaps by fully-convolutional networks (e.g., [5]) as an intermediate representation. These heatmaps are spatially discretized arrays (one for each joint), in which higher values indicate higher confidence that the particular joint is located at the corresponding position.

One line of 3D pose research builds on top of 2D heatmaps and infers the 3D pose from them by exemplar-matching [25], regression [8] or probabilistic inference [26]. One downside of such approaches is that the image content only indirectly influences the 3D estimation, as it acts on the result of the 2D estimation stage. Furthermore, 2D-to-3D lifting is performed in a numerical coordinate representation, which does not benefit from the built-in convolutional structure of CNNs.

Nibali et al. [12] predict three marginal heatmaps per body joint, for the XY, XZ and YZ planes, respectively. Pavlakos et al. [13] propose extending 2D heatmaps with a root-relative metric depth axis. One can obtain the 2D pixel positions and root-relative depths of the joints by finding maxima in the heatmaps.

One downside of heatmap representations has been the requirement of a dense output, which can become especially costly in 3D. The recently proposed soft-argmax [18], [19], [27] a.k.a. integral regression [14] method greatly alleviates this problem. As opposed to hard-argmax, which simply finds the location of the highest heatmap activation, soft-argmax is computed as the weighted average of all voxel grid coordinates, using softmaxed heatmap activations as the weights. For example, a low resolution heatmap can encode a joint position lying halfway between two bin centers by outputting 0.5 for both bins. By virtue of being differentiable, unlike hard-argmax, it also obviates the need for explicit heatmap-level supervision (e.g., voxel-wise cross-entropy). Instead, the loss can be computed (and its gradients back-propagated) from the coordinates yielded by soft-argmax.

Besides 2D heatmaps, Mehta et al. [9] estimate three further output channels per joint, the so-called location maps. These are read out at the position of the corresponding heatmap’s peak to obtain the X, Y and Z coordinates on a metric scale. Note how in their approach the final 3D joint coordinates are generated in the form of activation values (of the location maps at the heatmap peaks), as opposed to high-activation locations. We can thus think of it a conceptual hybrid of direct numerical coordinate regression and heatmap estimation. A downside of this method is that it requires high-resolution location maps and cannot benefit from the soft-argmax approach.

B. Scale and Distance Estimation

It is well-known that projecting a 3D world onto a 2D image plane results in ambiguity between size and distance (depth). However, the end goal for 3D scene understanding and 3D human pose estimation in particular is a metric-space output at the true scale. The ambiguity can only be resolved using semantic scale cues, i.e., prior knowledge of the usual size of humans and other objects appearing in the scene. Unfortunately, not all papers describe how this step is performed. Some authors report their results assuming an already known ground-truth root joint distance and focal length [12], [14], [28], [29]. A simple anthropometric approach is used by Pavlakos et al. [13] Given 2D pixel positions and root relative depth estimates from volumetric heatmaps, they optimize the absolute person distance such that the back-projected skeleton’s bone lengths match the average over the training set in a least squares sense, assuming a full perspective model. We use this scale recovery approach as our main baseline comparison throughout the paper, described in more detail in Section V.

Sun et al. [17] employ a similar idea, but use the overall skeleton length and a weak perspective model instead. Methods that are not based on volumetric heatmaps [30], [31] can directly predict the metric-scale numerical coordinates. Some recent works have shown that direct regression of person height from an image is a challenging task [32], [33].

While monocular 3D pose estimation methods are typically only evaluated in a root-relative manner, some works have also explicitly tackled the absolute (non-root-relative) pose estimation task, where every joint position is predicted within the 3D camera coordinate frame. This is closely related to the above-discussed metric-scale prediction: if both the image-space pose and the metric-scale root-relative pose are known, one can reconstruct the absolute distance (assuming a calibrated camera). Mehta et al. [34] and Dabral et al. [35] reconstruct the root offset by assuming a weak perspective model. Mehta et al. [36] assume the foot touches the known ground plane in the first frame. Moon et al. [37] predict the metric area of the human bounding box as a numerical value via a separate deep network (RootNet), besides their root-relative 2.5D PoseNet. In contrast to Moon et al., we estimate the scaled pose fully-convolutionally and do not require multiple separate backbones. In our earlier work [38], we estimate the distance directly from the image crop, but that does not generalize well to novel environments. Dabral et al. [35] propose to estimate the focal length jointly with the distance, implicitly relying on the perspective distortion of people far from the optical axis. As the authors note, this cannot work well when the camera is turned directly towards the target person. Véges and Lórizc [39] make use of a monocular depth prediction network pretrained on various indoor and outdoor datasets to help with absolute person distance estimation. Finally, some recent works also consider the depth relations among people: Jiang et al. [40] optimize the depth
ordering by occlusion cues, while Fieraru et al. [41] explicitly localize contact points between people to help with coherent reconstruction. In contrast, we perform our estimation for each person independently.

C. Truncated Pose Estimation

Single-person 3D pose estimation benchmarks, such as Human3.6M [42], [43], assume that the whole person is visible in the input image. In practical applications, however, bounding boxes are obtained using imperfect detectors, which can result in body truncation, especially in high-occlusion scenes. A possible remedy could be extending the detection crops by amodal completion [44], but this would result in a loss of image resolution. Generally, pose estimation performance under truncation has not been studied extensively in the literature. Recent work by Park et al. [45] uses cropping data augmentation to improve 2D pose estimation. Vosoughi and Amer create randomly truncated crops from Human3.6M images, and show that current methods perform poorly on truncated person images, even when only considering the present (within-boundary) joints [46]. They tackle the problem using direct numerical coordinate regression, similar to early 2D pose estimation methods [24]. We show that our approach performs significantly better in the truncated setting.

III. SINGLE-PERSON ROOT-RELATIVE APPROACH

In this section we present our proposed approach for metric-scale root-relative 3D human pose estimation. The input is an RGB image crop \( I \in \mathbb{R}^{w \times h \times 3} \) depicting a person. The desired output is a 3D skeleton, consisting of \( J \) joint coordinates \( \{ (\Delta X_j, \Delta Y_j, \Delta Z_j) \} \) in millimeters, up to arbitrary translation (hence the \( \Delta \) symbols).

A. Metric-Space Volumetric Heatmap Representation

As is common in heatmap-based approaches, we apply a fully-convolutional backbone network, with effective stride \( s \) to produce an array with \( d \cdot J \) spatial output channels. Here \( d \) is the number of discretization bins along the depth axis of the prediction volume. We then split the array along the channel axis into \( J \) volumes, each of shape \( (w/s) \times (h/s) \times d \). 3D spatial softmax is applied over each of them, resulting in volumetric heatmap activations \( V^{(i)} \in \mathbb{R}^{(w/s) \times (h/s) \times d} \). Up to this point the process is similar to other volumetric heatmap approaches [13], [14].

The difference lies in how the heatmap axes are interpreted to yield metric-scale coordinates. In particular, the 3D joint coordinates are decoded using soft-argmax with fixed scaling factors:

\[
\begin{bmatrix}
\Delta X_j \\
\Delta Y_j \\
\Delta Z_j
\end{bmatrix} = \sum_{p,q,r} V_{p,q,r}^{(i)} \cdot \begin{bmatrix}
p \cdot s/w \cdot W \\
q \cdot s/h \cdot H \\
r \cdot 1/d \cdot D
\end{bmatrix},
\]

(1)

where the \( p, q, r \) are 0-based integer indices into the volumetric heatmap array and \( W, H, D \) are the fixed metric width, height and depth extents of the full prediction volume. We set these extents as 2.2 meters in our work, which allows capturing people of usual height even when stretched out. Depending on striding logic (see Section III-C), Eq. (1) needs to be adjusted slightly, e.g., the volume size may change with denser striding (Fig. 3). The final root-relative prediction is obtained by subtracting the predicted root coordinates from all joint positions. Supervision is applied on these root-relative coordinates. This means that the position of the root joint prediction within the volume is not explicitly supervised and the network can place the skeleton anywhere within the prediction volume. The gradients are backpropagated through the root-joint-subtraction operation. No camera calibration-based back-projection, nor bone or skeleton size-based rescaling is needed for this root-relative prediction. The network is trained to perform these operations implicitly within the backbone.

B. Architecture

In contrast to prior work that employs decoders with upsampling layers and multiple refinement stages, we show that the task can be tackled in a significantly simpler fashion. Indeed, we apply the widely used ResNet-50 [47] backbone to directly predict spatial heatmaps, without any additional learnable layers, such as transposed convolutions. By default, ResNet-50 has an effective stride of 32, resulting in heatmaps of spatial size \( 8 \times 8 \) from the input image of size \( 256 \times 256 \) during training. The depth of the volumetric heatmap is set to 8. When testing on single-person datasets, we apply the trained network with an effective stride of 4, to obtain heatmaps with spatial size 64, which is the typical size used in prior work [13], [14].

This is called dense prediction and is commonly used in image segmentation [48]. In this technique, striding is removed from a given number of convolutional layers and the dilation rate of subsequent convolutions is increased correspondingly. As we will see, dense prediction increases the compute requirements but also improves accuracy, while still allowing real-time execution.

C. Centered Striding

When changing striding density at test time compared to training time, it is important to consider how the distribution of heatmap receptive field centers is affected. The left side of Fig. 3 shows a \( 256 \times 256 \) image processed with training stride 32 (+) and test stride 16 (×). The coverage changes significantly between training and test and is not symmetric over the image. While not an issue for pixel-labeling tasks, soft-argmax is a weighted vote-averaging scheme and introducing new voting positions in an uneven manner skews the prediction result. To tackle this issue, we propose centered striding, where the striding logic in the last convolutional layer of the backbone is “reversed”, such that it outputs the bottom right result per each \( 2 \times 2 \) block. The result is a more evenly distributed coverage over the image, with each original sampling position replaced with four new ones equally spaced around it. This benefit is evaluated in Section VII.

D. Scale and Translation Anagistic 2D Loss

Similar to recent approaches [10], [14], [16], we train simultaneously on 3D-labeled data from motion capture studies and 2D-labeled, in-the-wild data from the MPII dataset [49], to
Focusing on simplicity, no learnable parameters are introduced outside the standard backbone, except for a single 1×1 convolution. Optionally, if absolute (non-root-relative) pose estimation is required, our MeTRAbs extension also estimates classic 2D image-space heatmaps via another 1×1 convolutional head. We then reconstruct the absolute pose through a differentiable reconstruction module. This is based on a linear least squares formulation derived from the pinhole camera model. Supervision is applied both at the outputs of the individual prediction heads and at the final combined output. (*For 2D-labeled examples, the root-relative loss is replaced by a scale and translation-invariant 2D loss and the absolute 3D loss is not used.)

E. Truncated Pose Estimation

Our metric-space heatmap representation decouples the image boundary from the heatmap boundary. This enables the prediction of joint locations outside the image frame without additional design effort, the network is simply trained to output complete poses at a metric scale, regardless of how the input image is scaled or cropped. To evaluate this aspect, we follow Vosoughi and Amer [46] by randomly cropping H3.6M inputs, keeping at least 1/4 of the area of the person bounding square. Examples of such crops are in the second row of Fig. 7. We consider two scenarios. In the first one, the above described sampling of truncated crops is only performed at test time. In the second case, such crops are used for training as well.

F. Training

Prior work has shown that the $L^1$ loss is preferable in soft-argmax-based pose estimation [14]. To balance the losses computed on 3D and 2D-annotated examples, we use a fixed weighting factor $\lambda = 0.1$ tuned on a separate validation set of Human3.6M, yielding the overall loss as

$$L = L_{ann3D} + \lambda L_{ann2D}. \quad (2)$$

We initialize the network with ImageNet-pretrained weights and use the Adam optimizer with weight decay [51] and a batch size of 64. We decay the learning rate exponentially by an overall factor of 100, in two parts: from $10^{-4}$ to $3.33 \times 10^{-5}$ over 25 epochs and from $3.33 \times 10^{-6}$ to $10^{-6}$ in 2 final cooldown epochs.

As usual in deep learning, several sources of randomness influence the exact results of an experiment: random weight initialization, data shuffling, data augmentation and hardware-level non-determinism of execution order. We control these (except the last) by consistently seeding the random number generators. To distinguish random fluctuations from algorithmic differences, we repeat our experiments with 5 different...
As described above, our network is trained to output complete skeletons at the same metric scale, regardless of image zooming and truncation. To gain more intuition, we illustrate in Fig. 4 how this fully-convolutional model is able to achieve approximately image-scale- and truncation-invariant predictions. In particular, we can see that the soft-argmax output is not necessarily in the middle of the heatmap’s most prominent peak. As soft-argmax yields the heatmap’s center of mass, even distant heatmap values have an influence. Intuitively, this allows the network to move the prediction result towards different heatmap locations by adding counter-balancing correction weights, for example at the image sides or at the person center. Regarding truncation, the last row shows that the model can infer that the arms must lie above waist level, as there is no visual evidence of them in the image. To understand how a fully-convolutional network can “know” where the truncation happens, we refer to Islam et al.’s paper [52], showing that even fully-convolutional networks can encode positional information as a result of the zero-padding within convolutional layers. This means that the location of the top image border can be used as a cue for the network to shift the full skeleton downwards inside the heatmap volume, such that it fits. Note that the network is free to place the skeleton anywhere within the volume, since the root prediction is subtracted before computing the root-relative loss.

IV. Multi-Person Absolute Pose Approach

In this section, we propose MeTRAbs, a combination of MeTRo 3D heatmap estimation presented in Section III with traditional 2D pose heatmaps in a single end-to-end trained network for absolute 3D pose estimation. The main idea is that the MeTRo approach implicitly estimates the scale, which we then use to infer the distance. By applying this method within a top-down paradigm (detection, cropping, pose estimation), we obtain a fast and accurate way to tackle multi-person absolute 3D pose estimation.

Using our approach from Section III, we estimate a complete metric-scale pose \([(\Delta X_j, \Delta Y_j, \Delta Z_j)^T]_{j=1}^J\), up to translation (where \(J\) is the number of joints).

By additionally estimating the 2D, image-space pose \((x_j, y_j)^T\), we obtain all the necessary information to recover the absolute 3D pose in the (calibrated) camera coordinate system, as we will see in the following. For absolute pose estimation we assume known camera intrinsics, since monocular focal length estimation [44], [53] is a very challenging task (cf. the “dolly zoom” effect [54]). However, note that our method does not require the intrinsic calibration for root-relative estimation.

The absolute pose can be expressed as \([X_0 + \Delta X_j, Y_0 + \Delta Y_j, Z_0 + \Delta Z_j]^T\), with \((X_0, Y_0, Z_0)\) being the absolute pose offset, which we aim to recover. For this, we first calculate the normalized image coordinates as \((\bar{x}_j, \bar{y}_j)^T = K^{-1}(x_j, y_j)^T\), where \(K\) is the intrinsic matrix.

Mehta et al. [34] derive a formula to reconstruct the absolute root position using the weak perspective projection model. Véges and Lőrincz [39], while still operating in the weak perspective model, note that an approximation step involved in Mehta et al.’s algorithm leads to worse performance. Motivated by this, we derive a reconstruction method under the full perspective pinhole camera model and extensively compare it with Mehta et al.’s weak perspective method. In a full perspective model, a perfect estimate would satisfy

\[
\begin{pmatrix}
\bar{x}_j \\
\bar{y}_j
\end{pmatrix}
= \begin{pmatrix}
X_0 + \Delta X_j / (Z_0 + \Delta Z_j) \\
Y_0 + \Delta Y_j / (Z_0 + \Delta Z_j)
\end{pmatrix},
\]

which can be rearranged to

\[
\begin{pmatrix}
X_0 - \bar{x}_j Z_0 \\
Y_0 - \bar{y}_j Z_0
\end{pmatrix}
= \begin{pmatrix}
\bar{x}_j \Delta Z_j - \Delta X_j \\
\bar{y}_j \Delta Z_j - \Delta Y_j
\end{pmatrix}.
\]

Considering all joints, we obtain 2J linear equations in the three variables \((X_0, Y_0, Z_0)\). Since \(\bar{x}, \bar{y}, X, Y\) and \(Z\) are estimates, the equation system is noisy and over-determined. Hence we opt to solve it by linear least squares, with TensorFlow’s differentiable solver based on Cholesky decomposition. This differentiability allows us to directly supervise
the network with a loss \( L_{\text{abs3D}} \) computed on the final absolute 3D pose output, which turns out to be crucial for accurate distance estimation.

For truncated images, Eq. (3) only holds for body joints inside the image frame, since the 2D heatmap method cannot estimate out-of-image joint locations. We therefore exclude joints from the optimization, which are predicted to lie closer to the image border than one stride length. After reconstructing the root joint position, we can obtain the absolute pose in two ways. Either as \((\Delta X_j + X_0, \Delta Y_j + Y_0, \Delta Z_j + Z_0)^T\) (adding the reconstructed offset to the 3D head’s root-relative output), or as \((\tilde{x}_j, \tilde{y}_j, 1)^T \cdot (\Delta Z_j + Z_0)\) (back-projecting the 2D head’s output). For joints that lie within the image, we use the latter option, while for truncated ones we use the former. Both the individual prediction heads and the final absolute output are supervised with the \( L^1 \) loss. As in the root-relative MeTRo network, we apply weak supervision from 2D-labeled data for MeTRABs as well, on both heads. Extending Eq. (2), the loss becomes

\[
L = L_{\text{ann3D}} + L_{\text{head3D}} + \lambda \left( L_{\text{ann2D}} + L_{\text{head2D}} \right),
\]

where we again set \( \lambda = 0.1 \).

We found that the absolute loss can introduce numerical instabilities very early during training, since at this point the two prediction heads do not yet produce sufficiently compatible outputs, making the reconstruction problem ill-conditioned. Hence, we only turn on the absolute loss after 5000 update steps.

In a multi-person scenario, inference speed becomes a priority, since the model is evaluated on each person detection separately. To retain real-time performance, we do not apply dense prediction with MeTRABs; the network is trained and tested with coarse, \(8 \times 8\) heatmaps.

V. 2.5D BASELINE

For comparison, we implement a 2.5D baseline derived from Pavlakos et al.’s work [13], which introduced volumetric heatmaps for 3D human pose estimation. Pavlakos et al. use a coarse-to-fine estimation scheme with a stacked hourglass architecture [5] and no soft-argmax. To make the baseline directly comparable to our results, we instead use the architecture depicted in Fig. 5. This baseline directly estimates 2.5D heatmaps through a \(1 \times 1\) convolution at the end of the backbone. We then use soft-argmax, and compute the \( L^1 \) loss on the resulting coordinates. This makes the baseline similar to Sun et al. [14], except Sun et al. used additional learned layers and did not perform scale recovery. As a test-time post-processing step, the baseline uses the bone-length optimization method from Pavlakos et al. [55] to recover the root joint depth, which we briefly reiterate here. Given an assumed value for the root joint depth \( Z_0 \) and known camera intrinsics, the 2.5D pose can be back-projected into metric space and each bone’s resulting length \( b_i(Z_0) \) can be calculated. The optimal \( Z_0 \) is then the one that minimizes the squared bone length discrepancy, as compared to the average training bone lengths \( t_i \):

\[
Z_0^* = \arg \min_{Z_0} \sum_{i \in \text{bones}} (b_i(Z_0) - t_i)^2,
\]

where we only use bones, whose both ends are predicted to lie within the image (further from the border than 1 stride length). This is a convex, nonlinear least-squares problem, and we solve it using the Levenberg-Marquardt algorithm initialized at \( Z_0 = 2 \) m. To reiterate, as in [13], the absolute pose is not supervised during the baseline’s training and the convex optimization of \( Z_0 \) is not backpropagated through, for simplicity. We note, however, that the recent development of differentiable convex optimization layers [56], [57] could, in principle, enable such a solution as well.

VI. DATASETS, PREPROCESSING, EVALUATION

We conduct our single-person experiments on the largest available benchmarks: Human3.6M (shortened as H3.6M) [42], [43] and MPI-INF-3DHP (3DHP) [34]. The extended approach described in Section IV is evaluated in a multi-person context by training on MuCo-3DHP (MuCo) and testing on MuPoTS-3D (MuPoTS).

H3.6M [42], [43] was captured with 4 cameras in a motion capture studio. Two evaluation protocols are in wide use. In Protocol 1, the training subjects are 1, 5, 6, 7, 8, while 9 and 11 are used for testing. Prediction and ground-truth are aligned at the root joint, but no Procrustes alignment is performed. In Protocol 2, subjects 1, 5, 6, 7, 8, 9 are used in training and 11 in evaluation, with Procrustes alignment between prediction and ground truth. Every 64th frame is evaluated. We use the provided bounding boxes. We downsample videos from 50 to 10 fps. To further reduce redundancy, training frames are only used if at least one body joint moves at least 100 mm since the previous kept frame.

MPII [49] is a 2D-labeled dataset with 25k training images. We use this dataset for weak supervision, following the idea of Zhou et al. [10]. Only arm and leg joints are used from MPII, as we found these to be the most consistently labeled across datasets. In single-person experiments we only use instances explicitly marked as “well-separated” from other people and
take the provided person centers and sizes as the center and side length of the bounding box. In multi-person experiments, we use all person instances and the boxes are obtained with YOLOv3 [58].

**3DHP** [34] shows 8 training subjects in a green-screen studio. Test frames come from 3 scenes, each with 2 subjects: green-screen studio, studio without green screen, and outdoor. The latter two make this benchmark more challenging than H3.6M. In this dataset, the hips are labeled closer to the legs than in MPII. Following [10], we move the hips towards the neck by a fifth of the pelvis-neck vector before comparing with MPII-annotated skeletons for 2D loss computation. 3DHP provides two ground truth variants: unnormalized metric-space poses and “universal” (height-normalized) ones. We evaluate on both. We use only the chest-height cameras as [34], and only examples where all joints are within the image. We generate 3DHP bounding boxes by combining the bounding box of labeled joints and the most confident person detection of YOLOv3. The same frame sampling strategy is used as described above for H3.6M. Since its publication, the official 3DHP ground truth has been changed twice, making not all published results comparable. In our experience the first update changes scores by 1-3%, while the second one only by 0.1%, which is within experimental fluctuation, making the two latest versions comparable.

**MuCo** [31] is a synthetically composited multi-person dataset, derived from 3DHP by pasting persons over each other based on their root joint depth order. As [39], we generate 150k training images, each with 4 people. We run YOLOv3 on these images to get realistic bounding boxes.

**MuPoTS** [31] is a mixed indoor and outdoor multi-person test set, compatible with MuCo, consisting of 20 sequences showing people performing various actions and interactions. Like 3DHP, MuPoTS also provides normalized and unnormalized skeletons.

We crop images to the person’s bounding square and resize it to 256 × 256 px. Perspective effects must be taken into account when centering the image on the subject as this induces an implicit camera rotation [34]. We compensate by transforming both the input image and the predicted pose according to the implied camera rotation. The indoor 3DHP and MuPoTS sequences are gamma-corrected with an exponent of 0.67.

We apply geometric augmentations (scaling, rotation, translation, horizontal flip) and color distortion (brightness, contrast, hue, saturation). For single-person datasets, synthetic occlusion is added with 70% probability, half of which are rectangles with uniform white noise as in [70], half are segmented non-person objects from Pascal VOC [71] as in [38], [72]. On MuCo, synthetic occlusion probability is reduced to 30% since some occlusion is already introduced from compositing person segments over each other. On 3DHP and MuCo, we also augment the background with 70% probability following [34]. Backgrounds are taken from INRIA Holidays [73], excluding person images. All evaluation is done on a single crop, with no test-time augmentation.

We use the standard evaluation measures. The main one on 3DHP and MuPoTS is the percentage of correct keypoints (PCK), i.e., the fraction of joints predicted within 150 mm of the ground truth. The AUC is the area under the PCK curve as the threshold ranges from 0 to 150 mm. The measure on H3.6M is the mean per joint position error (MPJPE). 3DHP and MuPoTS evaluate 14 joints, excluding the root, while H3.6M uses 17, including the root. The official MuPoTS evaluation script rescales the bone lengths of the prediction to match the ground truth bone lengths before computing metrics, leading to some confusion and inconsistency between reported results. In [31] rescaling was only used for evaluating LCR-Net [74], but it has since been adopted by other authors as well. For consistency and simplicity, we train MeTRAbs only with unnormalized skeletons. When evaluating on universal (normalized) skeletons, we use bone rescaling. On unnormalized skeletons, we do not use bone rescaling, in order to directly evaluate the raw metric-space outputs of the methods. Following Véges and Lőrincz [75], on MuPoTS we also evaluate absolute (i.e., non-root-relative) versions of these metrics, prefixed with “A-”, e.g., A-PCK. For absolute MPJPE, Véges and Lőrincz [39], [75] evaluate all 17 joints and 16 (no pelvis) for relative MPJPE (but use 14 for PCK and A-PCK). For consistency, we always use 14 joints on MuPoTS, except when marked otherwise.

**VII. Results**

On H3.6M without ground truth depth or scale information, we achieve 49.3 mm MPJPE, which is within the margin of error compared to the state-of-the-art by Xu et al. [65] (49.2 mm), while using a considerably simpler approach (see Table I). (In all tables, the number after “±” is the standard deviation of 5 repeated experiments with different random seeds, therefore the standard error of the mean is a fifth of this value.) Besides simplifying the prediction pipeline and allowing for truncation-robust prediction (see below), our metric heatmap representation also performs better than the 2.5D baseline with bone-length-based scale recovery under the exact same experimental conditions. Table VII shows that training data augmentations improve performance by a large margin. On Protocol 2 (Table II), the benefit of our method is masked by the use of Procrustes alignment, which explicitly ignores
the quality of scale recovery. It is therefore unsurprising that our method performs about equally well as the 2.5D variant. On 3DHP, our method outperforms prior work by a large margin, including ones trained on more datasets as well (Table III). Both with universal (height-normalized) skeletons and true metric-scale ones, the MeTRo representation outperforms the baseline on green-screen studio images, however, the outdoor scenes were recorded on an empty field without scale cues and the explicit bone-length-based scale recovery performs better there. Qualitative results are in Fig. 7.

We analyze scale recovery in more detail (Table IV). As expected, the idealized method with test-time access to the ground truth root joint depth performs best on both H3.6M and 3DHP. The proposed approach performs better than the
TABLE IV
Comparison With Baseline Methods of Scale Recovery, With or Without Access to Ground Truth Information. For Both Datasets, Metric-Scale Skeletons Are Used With the Same 17 Joints for Comparability. The First Two Comparison Methods Access the Ground Truth at Test Time.

|                | H3.6M | 3DHP |
|----------------|-------|------|
|                | PCK↑  | AUC↑ | MPJPE | PCK↑  | AUC↑ | MPJPE |
| 2.5D GT root depth | 96.6  | 68.8 | 49.0  | 90.8  | 56.1 | 74.2  |
| 2.5D GT bone length | 96.4  | 67.0 | 51.9  | 90.3  | 56.1 | 74.6  |
| 2.5D avg train bones | 96.6  | 68.1 | 50.2  | 89.6  | 52.1 | 80.6  |
| MeTRo (ours)     | 97.0  | 68.6 | 49.3  | 93.6  | 52.6 | 81.1  |

TABLE V
MPJPE Scores on H3.6M Under Truncation, Evaluating All or Only the Present Joints. (*Training was not performed with truncated crops.) Results of Other Methods Are Taken From [46].

|                |                |                |                |
|----------------|----------------|----------------|----------------|
|                | Mehta [9] Zhou [10] Vosoughi [46] MeTRo MeTRo |
| All joints    | 396.4          | 400.5          | 415.0          |
| Present joints| 338.0          | 332.5          | 317.6          |

TABLE VI
Test Speed (Crops Per Second, FPS) and Error (H3.6M MPJPE) Tradeoff With the Two Striding Variants From Fig. 3.

|                | Striding variant | Test stride |
|----------------|------------------|-------------|
|                |                  | 32          | 16          | 8           |
| Speed          | Normal strides   | 53.1        | 52.5        | 52.7        | 52.9        |
|                | Center-aligned   | 50.9        | 50.2        | 50.0        | 49.3        |
| (Crop per sec.)| No batching      | 160         | 150         | 105         | 38          |
|                | Batch size 8     | 511         | 475         | 292         | 92          |

TABLE VII
Augmentation Ablation on H36M.

|                | Geometry | Color | Occlusion | MPJPE |
|----------------|----------|-------|-----------|-------|
|                | √        | √     | √         | 58.0  |
|                | √        | √     | -         | 52.8  |
|                | √        | -     | √         | 49.3  |

TABLE VIII
Results on MuPoTS-3D. Detected, Unnormalized Poses, No Bone Rescaling. (*Re-Evaluated Public Results; Joint Count: 1’7, 1’6, Else 1’4)

|                | A-MPJPE | MPJPE | A-PCK↑ | PCK↑  | DetRate↑ |
|----------------|---------|-------|--------|-------|----------|
| 2.5D baseline  | 317.6   | 313.6 | 114.0  | 110.0 | 40.0±1.0 | 79.3±0.3 | 94.2±0.0 |
| MeTRabs        | 248.2   | 246.9 | 108.2  | 104.3 | 40.2±1.0 | 81.1±0.4 | 94.1±0.1 |
| w/o abs. loss  | 328.8   | 327.8 | 108.4  | 104.7 | 36.7±3.2 | 80.9±0.4 | 94.1±0.1 |

TABLE IX
Comparison of Weak (W) and Full (F) Perspective-Based Absolute Pose Reconstruction. The Two Letters Denote the Training and the Test Time Variant. (Unnormalized Skeletons, Without Bone Rescaling.)

|                |                |                |                |
|----------------|----------------|----------------|----------------|
|               | Persp. assumption | All annotations | Matched annotations |
|               | Training Test | A-PCK↑ | A-PCK↑ | A-PCK↑ |
| F             | F             | 37.2±1.7 | 76.2±0.5 | 39.3±1.7 | 79.9±0.5 |
| F             | W             | 39.4±1.6 | 76.2±0.5 | 41.6±1.6 | 80.0±0.5 |
| W             | F             | 38.6±1.6 | 77.1±0.4 | 37.6±1.8 | 80.2±0.5 |
| W             | W             | 38.1±1.6 | 77.2±0.4 | 40.2±1.9 | 81.1±0.4 |

TABLE X
Results on the 3DPW Challenge Dataset. (PA=Procrustes Analysis)

|                |                |                |                |
|----------------|----------------|----------------|----------------|
|               | MPJPE          | MPJPE-PAL      | PCK@50mm↑ AUC@200mm↑ |
| Known association to GT identity |
| Sun et al. [78] | 81.8          | 58.6          | 37.3          | 59.9          |
| Kissos et al. [79] | 83.2          | 59.7          | 42.4          | 62.3          |
| MeTRabs (ours)     | 68.8          | 49.7          | 48.8          | 66.8          |

Unknown association to GT identity

| MeTRabs (ours) | 85.1 | 56.7 | 45.8 | 63.2 |

2.5D baseline using average bone lengths on H3.6M and comparably on 3DHP. On H3.6M, MeTRo closes most of this scale recovery gap between the 2.5D average bone length baseline and the idealized variant using the true root. Interestingly, our approach outperforms even the 2.5D variant using ground truth bone lengths for each test frame. On 3DHP, MeTRo’s scale recovery performance is similar to the 2.5D baseline (equal PCK, better AUC, slightly worse MPJPE). Further, on this dataset, access to ground truth scale information provides a larger improvement than on H3.6M, highlighting the importance of testing on many subjects.

When tested on truncated crops, our method by far outperforms prior approaches (Table V). This is true even for our default training configuration, but performance improves substantially when training on truncated images as well. The method is robust to truncation of up to 7 or 8 joints (of the 17) before overall performance substantially degrades (Fig. 6). Given the obvious ambiguity introduced by truncation, it is noteworthy that even truncated joints can be estimated with as little as about 100 mm average error. Qualitative examples are in the second row of Fig. 7, showing that our method can handle strongly truncated cases as well.

On the multi-person MuPoTS-3D, our MeTRAbs approach yields state-of-the-art results. For height-normalized skeletons with bone rescaling (standard setting in prior work, Table XI), MeTRAbs outperforms the 2.5D baseline, and the baseline already reaches state-of-the-art results. Our method performs particularly well on test sequence 2, with heavy occlusions (e.g., Fig. 8, left). Removing the supervision with the absolute 3D loss worsens the absolute PCK of all poses from 38.4% to 35.0%. Surprisingly, the root-relative accuracy seems to improve when turning off the absolute loss. This is, however, hard to interpret, as Table XI shows an artificial evaluation setting with normalized-height skeletons and bone-rescaling, thereby removing some of the scale recovery aspect from
the evaluation. When evaluating on unnormalized skeletons without bone rescaling (Table VIII), it becomes clear that the absolute loss helps: absolute MPJPE improves from 328.8 mm to 248.2 mm, absolute PCK from 36.7% to 40.2%, with the root-relative metrics slightly improving as well. These are state-of-the-art results and improve over methods that are pre- or jointly trained on ground truth pixel-wise depth prediction datasets \[39\], \[75\]. Further, we can see that the absolute PCK score has high variance and therefore small differences are not necessarily meaningful. The standard deviation across 5 repeat experiments is around 1.4–3.2%, and the absolute results for individual test sequences varies strongly across different configurations. This is because the test examples are strongly correlated since they come from video sequences. Lastly, we note that the detection rate is essentially the same for all of our configurations (Table VIII), since we use the same detections, and the official evaluation script performs matching based on the 2D projection, which is very similar across these methods.

In Table IX we evaluate whether using the full perspective pinhole camera model in the absolute pose reconstruction module brings benefits. In the last two rows, the absolute loss is not used at training time. In the other cases we backpropagate the absolute loss gradients either through the weak or full perspective reconstruction method. We find that training on MuCo with the full perspective model improves the absolute results, but when testing on MuPoTS, it is better to use the weak model. This may be explained by the fact that people in the MuCo dataset are closer to the camera than in MuPoTS, resulting in stronger perspective effects in MuCo. To verify this, we computed the ratio of the farthest and closest joint’s depth max \(Z_j\) / min \(Z_j\) per pose. If this ratio is large, the weak perspective assumption is a bad approximation. The median and the 90th percentile of this ratio on MuCo is 1.22 and

| S1  | S2  | S3  | S4  | S5  | S6  | S7  | S8  | S9  | S10 | S11 | S12 | S13 | S14 | S15 | S16 | S17 | S18 | S19 | S20 | Avg† |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 67.7 | 49.8 | 53.4 | 59.1 | 67.5 | 22.8 | 43.7 | 49.9 | 31.1 | 78.1 | 50.2 | 51.0 | 51.6 | 49.3 | 56.2 | 66.5 | 62.9 | 62.9 | 66.1 | 59.1 | 53.8 |
| 61.0 | 60.9 | 64.4 | 63.0 | 69.1 | 30.3 | 65.0 | 59.6 | 61.1 | 83.9 | 68.0 | 68.6 | 62.3 | 59.2 | 70.1 | 80.0 | 79.6 | 67.3 | 66.6 | 67.2 | 66.0 |
| 87.3 | 61.9 | 67.9 | 74.6 | 78.8 | 48.9 | 58.3 | 59.7 | 78.1 | 89.5 | 69.2 | 73.8 | 66.2 | 56.0 | 74.1 | 82.1 | 78.1 | 72.6 | 73.1 | 61.0 | 70.6 |
| 94.4 | 77.5 | 79.0 | 81.9 | 88.3 | 77.6 | 81.9 | 75.7 | 90.2 | 99.4 | 79.2 | 79.1 | 75.0 | 72.7 | 81.1 | 89.9 | 86.1 | 81.6 | 81.7 | 76.2 | 81.8 |
| 85.1 | 67.9 | 73.5 | 74.2 | 74.2 | 52.5 | 67.7 | 63.6 | 56.3 | 77.8 | 76.4 | 70.1 | 65.3 | 51.7 | 69.5 | 87.0 | 82.1 | 80.3 | 78.5 | 70.5 | 71.3 |
| 89.5 | 75.9 | 85.2 | 83.9 | 85.0 | 73.4 | 85.8 | 66.7 | 59.1 | 60.5 | 74.6 | 81.9 | 67.0 | 55.9 | 80.8 | 90.6 | 90.0 | 81.1 | 86.8 | 78.2 |
| 89.7 | 65.4 | 67.8 | 73.3 | 77.4 | 47.8 | 67.4 | 63.1 | 78.1 | 85.1 | 75.6 | 73.1 | 65.0 | 59.2 | 74.1 | 84.6 | 87.8 | 73.0 | 78.3 | 71.2 | 72.1 |
| 88.8 | 70.5 | 62.1 | 65.5 | 70.2 | 57.8 | 63.4 | 55.1 | 79.3 | 70.4 | 72.3 | 65.4 | 55.3 | 65.2 | 81.3 | 77.2 | 74.1 | 66.2 | 57.9 | 74.2 | 67.6 |
| 93.0 | 76.4 | 88.6 | 85.2 | 86.3 | 75.7 | 84.3 | 67.9 | 84.3 | 93.4 | 81.6 | 89.8 | 77.3 | 83.8 | 91.0 | 86.1 | 84.8 | 78.1 | 72.1 | 82.3 | 80.1 |
| 93.8 | 80.8 | 89.3 | 87.0 | 86.6 | 74.5 | 83.7 | 66.2 | 85.0 | 92.9 | 80.4 | 89.6 | 77.1 | 68.7 | 86.3 | 92.0 | 86.4 | 74.4 | 73.7 | 82.4 | 80.3 |
| 93.8 | 80.8 | 89.3 | 87.0 | 86.6 | 74.5 | 83.7 | 66.2 | 85.0 | 92.9 | 80.4 | 89.6 | 77.1 | 68.7 | 86.3 | 92.0 | 86.4 | 74.4 | 73.7 | 82.4 | 80.3 |
| 93.0 | 80.1 | 89.2 | 85.8 | 90.1 | 76.9 | 88.6 | 75.6 | 84.3 | 93.4 | 85.4 | 90.6 | 83.4 | 80.9 | 83.8 | 93.0 | 86.6 | 89.3 | 85.0 | 90.8 | 86.3 |
| 93.8 | 94.4 | 90.0 | 87.6 | 90.5 | 75.7 | 88.1 | 74.9 | 85.0 | 92.9 | 84.7 | 90.4 | 83.3 | 82.2 | 86.3 | 93.9 | 87.1 | 88.9 | 85.2 | 91.3 | 88.6 |
| 93.8 | 94.4 | 90.0 | 87.6 | 90.5 | 75.7 | 88.1 | 74.9 | 85.0 | 92.9 | 84.7 | 90.4 | 83.3 | 82.2 | 86.3 | 93.9 | 87.1 | 88.9 | 85.2 | 91.3 | 88.6 |
| 93.0 | 80.1 | 89.2 | 85.8 | 90.1 | 76.9 | 88.6 | 75.6 | 84.3 | 93.4 | 85.4 | 90.6 | 83.4 | 80.9 | 83.8 | 93.0 | 86.6 | 89.3 | 85.0 | 90.8 | 86.3 |

### TABLE XI

**Comparison to Prior Work on the MuPoTS-3D Benchmark for Normalized Skeletons With Bone Rescaling to the Ground Truth Before Computing the Percentage of Correct Keypoints (PCK).**

(For the direct evaluation of the Metric-Space Poses, See Table VIII)

The table below displays the comparison to prior work on the MuPoTS-3D benchmark for normalized skeletons with bone rescaling to the ground truth before computing the percentage of correct keypoints (PCK). The table compares our method (METRO) with other methods listed in Table VIII. The results are presented in terms of root-relative and absolute PCK metrics for different evaluation settings.

| Evaluation Setting | METRO & Absolute Loss | METRO & w/o Absolute Loss |
|--------------------|-----------------------|---------------------------|
| 2D baseline        | 93.0                  | 94.0                      |
| MeTRABS            | 93.8                  | 94.0                      |
| w/o absolute loss  | 93.0                  | 94.0                      |

### Root-relative PCK for all annotated poses

| Evaluation Setting | METRO & Absolute Loss | METRO & w/o Absolute Loss |
|--------------------|-----------------------|---------------------------|
| 2D baseline        | 93.0                  | 94.0                      |
| MeTRABS            | 93.8                  | 94.0                      |
| w/o absolute loss  | 93.0                  | 94.0                      |
Fig. 7. Qualitative results on various datasets. Predictions are shown in color, ground truth in gray (except for MPII, where it is unavailable). Green spheres mark predictions within 150 mm of the ground truth, red cubes beyond that threshold. Note that our method performs well on truncated (partial body) images as well (second row).

Fig. 8. Qualitative results on MuPoTS-3D (prediction in blue-yellow, ground truth in red).

1.41, while on MuPoTS it is only 1.16 and 1.26, respectively. This confirms that perspective effects are stronger in MuCo.

Another possible reason is that the model may output slightly perspective distorted results in the metric 3D head, which are better handled by the weak-perspective model in the next step, as opposed to training time, when the network learns to output the correct metric, perspective-undistorted pose, for which the full perspective model works better afterwards. Nevertheless, as there is no clear overall winner between the weak and full perspective models, and changing the method across training and test is clearly not desirable, we use the more commonly applied weak perspective method for all other experiments.

A. Inference Speed

Our method is capable of real-time inference. The root-relative architecture can process 511 crops per second on an RTX 2080 Ti desktop GPU when operating on batches of 8 crops at stride 32 (Table VI). Varying the heatmap resolution using dense prediction provides diminishing returns in accuracy (Table VI), showing that soft-argmax can cope with heatmaps of very coarse resolution. By gathering all person instances of a frame in a batch, MeTRAbs can process 128, 118, 98, 67, 41 frames per second for 1, 2, 4, 8 and 16 people per frame, respectively. The above calculations assume the image crops are available instantly and the time cost of detection is excluded.

B. ECCV 3DPW Challenge

Finally, we note that our MeTRAbs method has won the 3D Poses in the Wild (3DPW) [21] challenge, organized as a workshop event at the European Conference on Computer Vision, 2020. Table X compares results using the 3DPW pro-
to follow. For this, we train our network on a combination of the H3.6M, MuCo, SURREAL [81], SAIL-VOS [82] and CMU-Panoptic [83] datasets. We use ResNet-101 as the backbone and additionally apply upper body crop (truncation) augmentation at training time and 5-crop averaging at test time. When identity tracking is needed, we perform frame-to-frame matching based on absolute pose distance. The listed methods are not directly comparable due to different training data. Even with this caveat, our top results show that our approach can scale with further training data and performs well even in challenging in-the-wild scenarios.

VIII. CONCLUSION

We proposed metric-scale truncation-robust (MeTRo) volumetric heatmaps for the tasks of root-relative and absolute 3D human pose estimation. These heatmaps directly represent the metric space around the person instead of being tied to the image space and can be predicted with any standard fully-convolutional network. With a modified weak supervision scheme for 2D labels, careful stride alignment considerations and strong data augmentation, we achieved state-of-the-art results on two important single-person benchmarks: Human3.6M and MPI-INF-3DHP. We showed that our approach can implicitly discover scale cues from the data, given its superior performance compared to a previous, fixed bone length based heuristic on most test scenarios. Future research should consider possibilities for learning similar scale cues from large-scale outdoor data as well. Another interesting future direction can be the evaluation on people with widely differing heights, if such data becomes available on a large scale. Further, we demonstrated the second benefit of the MeTRo representation, the prediction (“hallucination”) of complete skeletons even when only a part of the body is contained in the image. For the multi-person absolute 3D pose estimation scenario, we developed a combination of MeTRo heatmaps with 2D heatmap prediction, referred to as MeTRAbs. We saw the importance of supervising the absolute pose prediction end-to-end by employing a differentiable combination of 2D and root-relative 3D poses. For this we tested two alternatives, based on weak and full perspective geometry, but neither performed clearly better than the other in our experiments. Applying MeTRAbs in the top-down multi-person paradigm, we achieved state-of-the-art results on the challenging MuPoTS-3D dataset while keeping the method real-time capable. Overall we can conclude that heatmap estimation is a versatile paradigm and it is possible to tackle absolute 3D human pose estimation through exclusively estimating heatmaps and encoding all quantities such as coordinates or sizes as activation locations, instead of as activation values.

ACKNOWLEDGMENT

The authors thank Yinglung Liu for help in evaluating on MuPoTS-3D. Compute resources were granted by RWTH Aachen University under Project rwth0479.

REFERENCES

[1] C. Zimmermann, T. Welschheold, C. Dornhege, W. Burgard, and T. Brox, “3D human pose estimation in RGBD images for robotic task learning,” in Proc. ICRA, 2018, pp. 1986–1992.
[2] T. Allieck, M. Magazin, X. Xu, C. Theobalt, and G. Pons-Moll, “Video based reconstruction of 3D people models,” in Proc. CVPR, 2018, pp. 8337–8397.
[3] V. Srivastav, T. Iwamoto, M. Kadkhodamohammadi, M. Mathelin, A. Gangi, and N. Padoy, “MVOR: A multi-view RGB-D operating room dataset for 2D and 3D human pose estimation,” 2018. [Online]. Available: arxiv:1808.08180.
[4] A. Neverova, R. A. Größer, and I. Kokkinos, “Dense pose transfer,” in Proc. ECCV, 2018, pp. 128–143.
[5] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in Proc. ECCV, 2016, pp. 483–499.
[6] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, “Learning feature pyramids for human pose estimation,” in Proc. CVPR, 2017, pp. 1290–1299.
[7] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, “Multi-scale structure-aware network for human pose estimation,” in Proc. ECCV, 2018, pp. 731–746.
[8] J. Martinez, R. Hossain, J. Romero, and J. J. Little, “A simple yet effective baseline for 3D human pose estimation,” in Proc. ICCV, 2017, pp. 2659–2668.
[9] D. Mehta et al., “VNeCT: Real-time 3D human pose estimation with a single RGB camera,” ACM Trans. Graphics, vol. 36, no. 4, 2017, Art. no. 44. [Online]. Available: https://dl.acm.org/doi/10.1145/3072959.3073596.
[10] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, “Towards 3D human pose estimation in the wild: A weakly-supervised approach,” in Proc. ICCV, 2017, pp. 398–407.
[11] T. Welschheold, C. Dornhege, W. Burgard, and T. Brox, “3D human pose estimation = 2D pose estimation + matching,” in Proc. ICCV, 2017, pp. 1263–1272.
[12] X. Sun, B. Xiao, S. Liang, and Y. Wei, “Integral human pose regression,” in Proc. ECCV, 2018, pp. 536–553.
[13] U. Iqbal, P. Molchanov, T. Breuel, J. Gall, and J. Kautz, “Hand pose estimation via latent 2.5D heatmap regression,” in Proc. ECCV, 2018, pp. 125–143.
[14] D. C. Luvizon, D. Picard, and H. Tabia, “2D/3D pose estimation and perspective reconstruction of 3D people models,” in Proc. CVPR, 2018, pp. 5137–5146.
[15] S. Levine, C. Finn, T. Darrell, and P. Abbeel, “End-to-end training of deep visual motor policies,” J. Mach. Learn. Res., vol. 17, no. 1, pp. 1–40, 2016.
[16] A. Nibali, Z. He, S. Morgan, and L. Prendergast, “3D human pose estimation with 2D marginal heatmaps,” in Proc. WACV, 2019, pp. 1477–1485.
[17] X. Sun, C. Li, and S. Lin, “An integral pose regression system for the ECCV2018 PoseTrack Challenge,” 2018. [Online]. Available: arxiv:1809.06079.
[18] Y. Chen, Y. Tian, and M. He, “Monocular human pose estimation: A survey of deep learning-based methods,” Comput. Vis. Image Understanding, vol. 192, Mar. 2020, Art. no. 102897.
[19] A. Tošhev and C. Szegedy, “DeepPose: Human pose estimation via deep neural networks,” in Proc. CVPR, 2014, pp. 1653–1660.
[20] C.-H. Chen and D. Ramanan, “3D human pose estimation = 2D pose estimation + matching,” in Proc. CVPR, 2017, pp. 5759–5767.
[21] D. Tome, C. Russell, and L. Agapito, “Lifting from the deep: Convolutional 3D pose estimation from a single image,” in Proc. CVPR, 2017, pp. 5689–5698.
István Sárándi (Student Member, IEEE) received the B.Sc. degree in computer engineering from the Budapest University of Technology and Economics, and the M.Sc. degree in computer science from RWTH Aachen University. He is currently pursuing the Ph.D. degree with the Computer Vision Group, RWTH Aachen University, supervised by Prof. Dr. Bastian Leibe. His research focuses on visual human analysis with an emphasis on 3D body pose for robotics applications. Methods from his first-author publications achieved First Place both in the 2018 ECCV PoseTrack Challenge on 3D human pose estimation and in the 2020 ECCV 3D Poses in the Wild Challenge.

Timm Linder (Member, IEEE) received the Ph.D. degree from the University of Freiburg, Germany, in 2020, thesis on multimodal human detection, tracking and analysis for robots in crowded environments. Since 2016, he has been a Research Scientist with focus on autonomous systems and robot perception with Bosch Corporate Research. He has coauthored peer-reviewed publications at major international conferences and journals, served on different program committees in robotics and AI. His research interests include computer vision, in particular human detection, tracking and pose estimation, as well as 3D scene generation and sim-to-real transfer. He received an Outstanding Reviewer Award at ICRA in 2019.

Kai Oliver Arras (Member, IEEE) received the Ph.D. degree from EPFL. He is the Head of Robotics Research and a Chief Expert in robotics with Robert Bosch GmbH. Until 2015, he was an Assistant Professor for social robotics, and HRI with the University of Freiburg where he was awarded a DFG Junior Research Group Leader Grant. He was a Postdoctoral Researcher with KTH Stockholm and the University of Freiburg. He published around 120 peer-reviewed papers, articles, editorials and book chapters on robot navigation, perception, planning, and system integration. He was a Member of various program committees in robotics, AI, HRI, and computer vision.

Bastian Leibe (Member, IEEE) received the M.S. degree in computer science from the Georgia Institute of Technology in 1999, the Diploma degree in computer science from the University of Stuttgart in 2001, and the Ph.D. degree in computer science from ETH Zurich in 2004. He is a Full Professor of Computer Science with RWTH Aachen University, Germany, where he leads the Computer Vision Group. He has published over 130 articles in peer-reviewed journals and conferences. His main research interests are in computer vision and machine learning for dynamic visual scene understanding, encompassing object recognition, tracking, segmentation, and 3D reconstruction. Over the years, he has received several awards for his research work, including the CVPR Best Paper Award in 2007, the DAGM Olympus Prize in 2008, and the U.V. Helava Award in 2012. He was awarded an European Research Council (ERC) Starting Grant in 2012, and an ERC Consolidator Grant in 2017. He has been a Program Chair for ECCV 2016 and an Area Chair and Program Committee Member for all major computer vision conferences.