

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

István Sárándi, *Student Member, IEEE*, Timm Linder, *Member, IEEE*,  
 Kai O. Arras, *Member, IEEE*, 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 (*MeTRo*) 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 our representation, we present a differentiable combination of our 3D metric-scale heatmaps with 2D image-space ones to estimate absolute 3D pose (our *MeTRAbs* 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.<sup>1</sup>

**Index Terms**—3D human pose estimation, absolute human pose, scale estimation, truncation

## 1 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

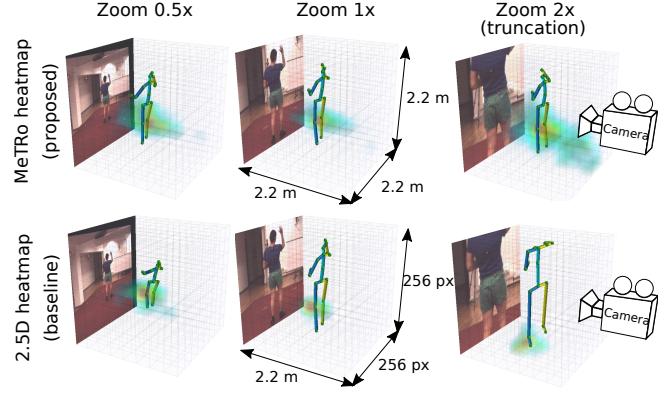


Fig. 1. By defining heatmaps in the 3D metric space around the person (top row) we directly estimate scale-correct complete poses. This is in contrast to prior work (bottom row) that defines the X and Y heatmap axes in image space and requires further post-processing to obtain a metric-scale skeleton. 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.

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

• I. Sárándi and B. Leibe are with RWTH Aachen University, Germany. Email: {Sarandi, Leibe}@vision.rwth-aachen.de

• T. Linder and K. O. Arras are with Robert Bosch GmbH, Renningen, Germany. Email: {Timm.Linder, KaiOliver.Arras}@de.bosch.com

© 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

<sup>1</sup> <https://vision.rwth-aachen.de/metraabs>

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 paper 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 paper 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.

## 2 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].

### 2.1 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 as 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.

## 2.2 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 Sec. 5. 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 *et al.* [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.

## 2.3 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

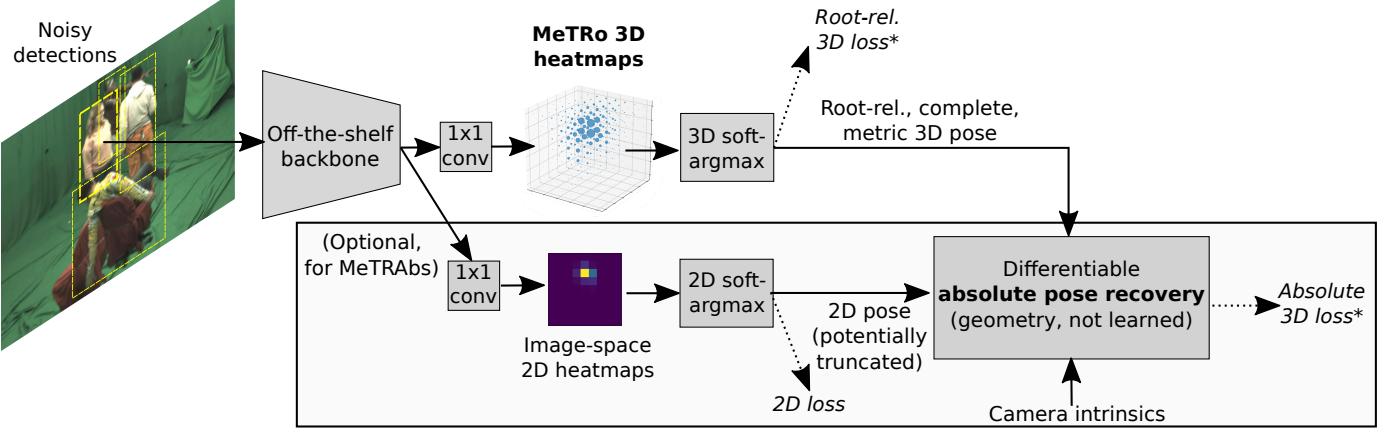


Fig. 2. Overview of our approach. We predict volumetric heatmaps using an off-the-shelf fully-convolutional backbone. Applying soft-argmax on these heatmaps and scaling by an image-independent constant factor yields joint coordinates in metric space up to translation. We minimize the root-relative  $L^1$  loss. Focusing on simplicity, no learnable parameters are introduced outside the standard backbone, except for a single  $1 \times 1$  convolution. Optionally, if absolute (non-root-relative) pose estimation is required, our *MeTRAbs* extension also estimates classic 2D image-space heatmaps via another  $1 \times 1$  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.)

literature. Recent work by Park *et al.* [45] uses cropping data augmentation to improve 2D pose estimation. Vosoughi *et al.* 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.

### 3 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  $\{(\Delta X_j, \Delta Y_j, \Delta Z_j)^T\}_{j=1}^J$  in millimeters, up to arbitrary translation (hence the  $\Delta$  symbols).

#### 3.1 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  $V^{(j)} \in \mathbb{R}^{(w/s) \times (h/s) \times d}$ . 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:

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

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 Sec. 3.3), 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.

#### 3.2 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,

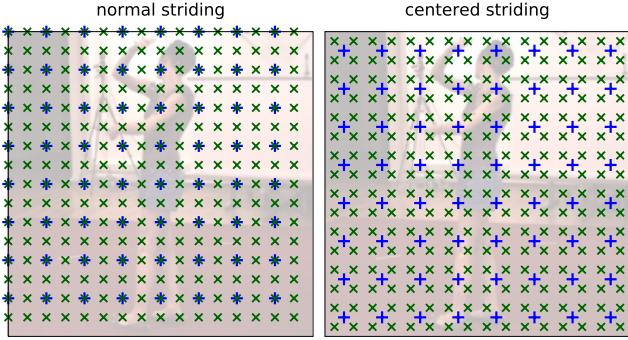


Fig. 3. Receptive field centers of the output neurons in a strided FCN operating on a  $256 \times 256$  image (+: stride 32,  $\times$ : stride 16). *left:* normal striding logic, where the top left result is kept per  $2 \times 2$  block. Note that denser striding skews the sample density towards the bottom and right in the border areas. *right:* by reversing the stride logic in the last strided layer (*i.e.*, bottom right result taken, instead of top left), the samples are centered and the increased striding density is distributed evenly.

dense prediction increases the compute requirements but also improves accuracy, while still allowing real-time execution.

### 3.3 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 ( $\times$ ). 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 Sec. 7.

### 3.4 Scale and Translation Agnostic 2D Loss

Similar to recent approaches [10], [14], [16], we train simultaneously on 3D-labeled data from motion capture studios and 2D-labeled, in-the-wild data from the MPII dataset [49], to incorporate more appearance variation in the training process. Half of each mini-batch is filled with examples of either kind. Supervision via 2D labels is straightforward when using 2.5D heatmaps, as the  $X$  and  $Y$  heatmap axes correspond to the space in which the 2D labels are defined. However, since our prediction volume is defined on a metric scale and is not aligned with image space, we propose a 2D loss computation method that is invariant to prediction scale and translation. To this end, we first orthographically project the predicted 3D skeleton onto the image plane by discarding the  $Z$  coordinate. Then we align the projected prediction to the 2D pixel-scale ground truth by translation and uniform scaling to the least-squares optimal fit before computing the loss. This alignment layer is differentiable and gradients can be backpropagated through it. We note that a

similar scale-invariant loss has been used by Rhodin *et al.* to enforce multi-view consistency of 3D poses [50].

### 3.5 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 *et al.* [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.

### 3.6 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

$$\mathcal{L} = \mathcal{L}_{\text{ann3D}} + \lambda \mathcal{L}_{\text{ann2D}}. \quad (2)$$

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 seeds and report the mean and standard deviation of the evaluation metrics.

### 3.7 Intuition

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

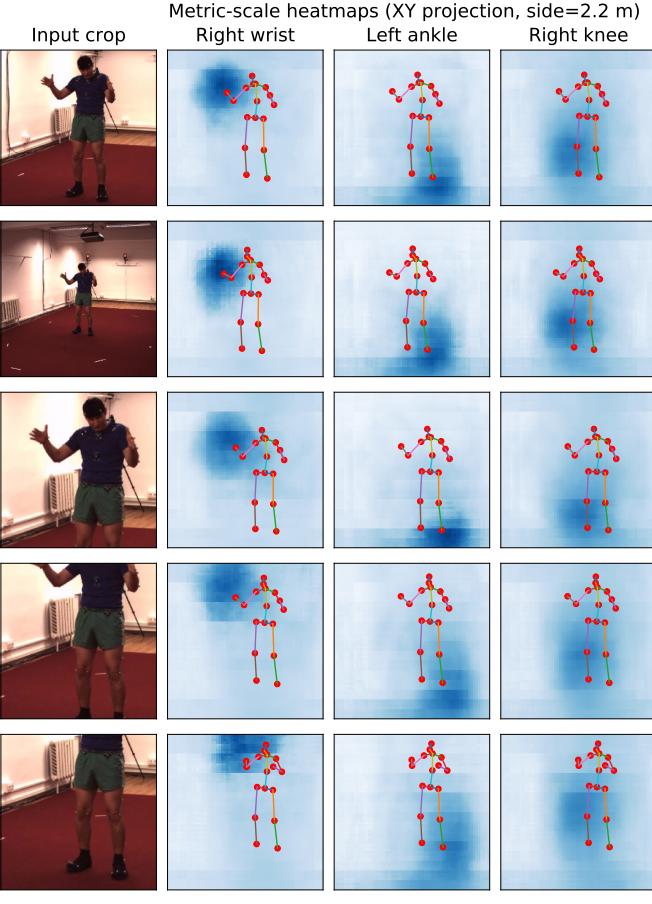


Fig. 4. A closer look at scale and truncation robustness. We plot the projected *metric-scale* heatmaps for 3 joints with the full soft-argmax skeleton for reference. The predicted skeleton is approximately invariant to change in scale and truncation. Since the metric size of the person does not change with image scaling, the backbone learns to output heatmaps with a similar center of mass, regardless of image scale. Note that the heatmaps do not align with image space and this is intended by design. (The broad peaks are a result of training the model at low,  $8 \times 8$  heatmap resolution.)

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-paddings 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.

#### 4 MULTI-PERSON ABSOLUTE POSE APPROACH

In this section, we propose MeTRAbs, a combination of MeTRo 3D heatmap estimation presented in Sec. 3 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 Sec. 3, 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  $\{(x_j, y_j)^T\}_{j=1}^J$ , 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 (*c.f.* 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  $\{(X_0 + \Delta X_j, Y_0 + \Delta Y_j, Z_0 + \Delta Z_j)^T\}_{j=1}^J$  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  $(\tilde{x}_j, \tilde{y}_j)^T = K^{-1}(x_j, y_j)^T$ , 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 *et al.* [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

$$\begin{bmatrix} \tilde{x}_j \\ \tilde{y}_j \end{bmatrix} = \begin{bmatrix} (X_0 + \Delta X_j)/(Z_0 + \Delta Z_j) \\ (Y_0 + \Delta Y_j)/(Z_0 + \Delta Z_j) \end{bmatrix}, \quad (3)$$

which can be rearranged to

$$\begin{bmatrix} X_0 - \tilde{x}_j Z_0 \\ Y_0 - \tilde{y}_j Z_0 \end{bmatrix} = \begin{bmatrix} \tilde{x}_j \Delta Z_j - \Delta X_j \\ \tilde{y}_j \Delta Z_j - \Delta Y_j \end{bmatrix}. \quad (4)$$

Considering all joints, we obtain  $2J$  linear equations in the three variables  $(X_0, Y_0, Z_0)$ . Since  $\tilde{x}, \tilde{y}, X, Y$  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  $\mathcal{L}^{\text{abs3D}}$  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

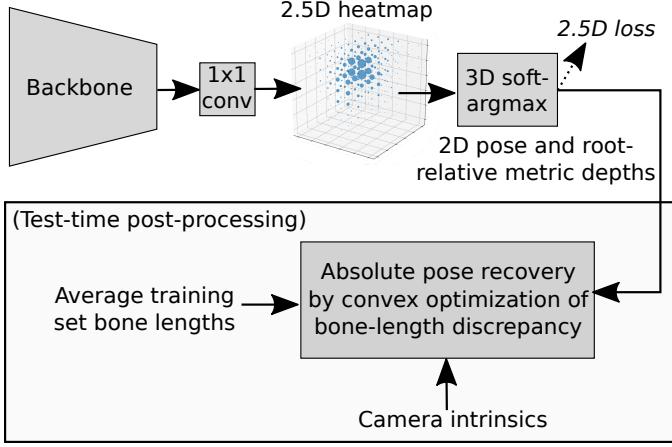


Fig. 5. Baseline architecture with 2.5D heatmaps for ablative comparison experiments.

$$\mathcal{L} = \mathcal{L}_{\text{ann3D}}^{\text{abs3D}} + \mathcal{L}_{\text{ann3D}}^{\text{head3D}} + \mathcal{L}_{\text{ann3D}}^{\text{head2D}} + \lambda (\mathcal{L}_{\text{ann2D}}^{\text{head2D}} + \mathcal{L}_{\text{ann2D}}^{\text{head3D}}), \quad (5)$$

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.

## 5 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  $t_i$ :

$$Z_0^* = \arg \min_{Z_0} \sum_{i \in \text{bones}} (b_i(Z_0) - t_i)^2, \quad (6)$$

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.

## 6 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 Sec. 4 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 64<sup>th</sup> 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  $256 \times 256$  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 *et al.* [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 *et al.* [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.

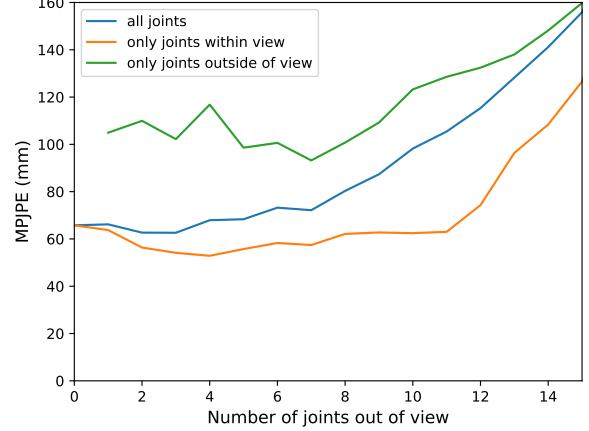


Fig. 6. Analysis of robustness to truncation on H3.6M. Average performance remains relatively stable up to 7 truncated joints.

## 7 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 Tab. 1). (In all tables, the number after “ $\pm$ ” 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. Tab. 7 shows that training data augmentations improve performance by a large margin. On Protocol 2 (Tab. 2), 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 (Tab. 3). 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 (Tab. 4). 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 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

TABLE 1

Evaluation on H3.6M Protocol 1 (subjects 9 and 11), using mean per joint position error (MPJPE) without Procrustes alignment.  
All methods use extra 2D-labeled pose data in training.

	Dir.	Dis.	Eat	Gre.	Phn.	Pose	Pur.	Sit	SitD	Sm.	Pht.	Wait	Walk	WD	WT	Avg ↓
<i>Methods using ground-truth scale or depth information at test time</i>																
Sun <i>et al.</i> [59]	52.8	54.8	54.2	54.3	61.8	53.1	53.6	71.7	86.7	61.5	67.2	53.4	47.1	61.6	53.4	59.1
Nibali <i>et al.</i> [12]	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	57.0
Luvizon <i>et al.</i> [16]	51.5	53.4	49.0	52.5	53.9	50.3	54.4	63.6	73.5	55.3	61.9	50.1	<b>46.0</b>	60.2	51.0	55.1
Luvizon <i>et al.</i> [60]	43.7	48.8	45.6	<b>46.2</b>	49.3	43.5	46.0	56.8	67.8	50.5	57.9	<b>43.4</b>	40.5	53.2	45.6	49.5
Sun <i>et al.</i> [14]	47.5	47.7	49.5	50.2	51.4	43.8	46.4	58.9	65.7	49.4	55.8	47.8	38.9	<b>49.0</b>	43.8	49.6
Chen <i>et al.</i> [29]	45.3	49.8	46.1	49.6	<b>48.2</b>	<b>41.7</b>	47.4	<b>53.1</b>	<b>55.2</b>	<b>48.0</b>	57.7	45.6	40.8	52.4	45.2	<b>48.4</b>
<i>Methods using no ground truth scale or depth information at test time</i>																
Pavlakos <i>et al.</i> [13]	67.4	72.0	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Zhou <i>et al.</i> [10]	54.8	60.7	58.2	71.4	62.0	53.8	55.6	75.2	111.6	64.2	65.5	66.0	51.4	63.2	55.3	64.9
Martinez <i>et al.</i> [8]	51.8	56.2	58.1	59.0	69.5	55.2	58.1	74.0	94.6	62.3	78.4	59.1	49.5	65.1	52.4	62.9
Fang <i>et al.</i> [61]	50.1	54.3	57.0	57.1	66.6	53.4	55.7	72.8	88.6	60.3	73.3	57.7	47.5	62.7	50.6	60.4
Yang <i>et al.</i> [62]	51.5	58.9	50.4	57.0	62.1	49.8	52.7	69.2	85.2	57.4	65.4	58.4	43.6	60.1	47.7	58.6
Pavlakos <i>et al.</i> [63]	48.5	54.4	54.4	52.0	59.4	49.9	52.9	65.8	71.1	56.6	65.3	52.9	44.7	60.9	47.8	56.2
Liu <i>et al.</i> [64]	47.0	53.1	50.3	48.8	56.0	48.1	47.6	65.9	72.6	52.3	61.4	49.1	39.3	54.2	<b>40.6</b>	52.4
Xu <i>et al.</i> [65]	<b>40.6</b>	<b>47.1</b>	45.7	<b>46.6</b>	50.7	45.0	47.7	56.3	<b>63.9</b>	49.4	63.1	<b>46.5</b>	<b>38.1</b>	51.9	42.3	<b>49.2</b>
Sharma <i>et al.</i> [66]	48.6	54.5	54.2	55.7	62.6	50.5	54.3	70.0	78.3	58.1	72.0	55.4	45.2	61.4	49.7	58.0
Cai <i>et al.</i> [67]	46.5	48.8	47.6	50.9	52.9	48.3	45.8	59.2	64.4	51.2	61.3	48.4	39.2	53.5	41.2	50.6
2.5D baseline	45.1	50.4	45.4	47.8	<b>50.0</b>	<b>44.6</b>	49.8	59.0	69.4	49.4	56.5	48.0	39.6	<b>49.4</b>	45.0	50.2±0.3
<b>MeTRo (ours)</b>	46.3	48.3	<b>43.3</b>	48.2	50.2	45.1	<b>46.1</b>	<b>56.2</b>	66.8	<b>49.3</b>	<b>54.5</b>	46.7	40.1	49.6	46.2	<b>49.3±0.7</b>

TABLE 2

Comparison of MPJPE with prior work on H3.6M under Protocol 2 (test subject 11 with Procrustes alignment to the ground truth).

Nie [68]	Pavlakos [13]	Sun [59]	Martinez [8]	Sun [14]	Nibali [12]	Habibie [69]	Xu [65]	Chen [29]	2.5D baseline	MeTRo (ours)	
P-MPJPE	79.5	51.9	48.3	47.7	40.6	40.4	49.2	38.9	<b>33.7</b>	34.5±0.4	34.7±0.5

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 (Tab. 5). 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, Tab. 11), 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 Tab. 11 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 (Tab. 8), 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 as well 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 (Tab. 8), 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 9 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 back-propagate 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  $\max_j Z_j / \min_j Z_j$  per pose. If this ratio is large, the weak perspective assumption is a bad approximation. The median and the 90th percentile of this

TABLE 3

Comparison on MPI-INF-3DHP with prior methods. \* Evaluated with the first version of the dataset, with some annotation difference. Dashes (–) reflect a lack of published information. Superscripts indicate the training data (first characters of 3DHP, H36M, MPII, LSP and COCO).

	Stand/ walk	Exer- cise	Sit on chair	Cro./ reach	On floor	Sport	Misc.	Green screen	No gr.sc.	Out- door	Total
	PCK↑								PCK↑	AUC↑	MPJPE↓
<i>Universal, height-normalized skeletons (simplified scale recovery task)</i>											
Rogez <i>et al.</i> [74]*	70.5	56.3	58.5	69.4	39.6	57.7	57.6	–	–	–	59.7
Zhou <i>et al.</i> <sup>H+M</sup> [10]*	85.4	71.0	60.7	71.4	37.8	70.9	74.4	71.7	64.7	72.7	69.2
Zhou <i>et al.</i> <sup>H+M</sup> [76]	–	–	–	–	–	–	–	75.6	71.3	80.3	75.3
Mehta <i>et al.</i> <sup>3+M+L+H</sup> [9]*	87.7	77.4	74.7	72.9	51.3	83.3	80.1	–	–	–	76.6
Mehta <i>et al.</i> <sup>3+M+L+H</sup> [34]*	86.6	75.3	74.8	73.7	52.2	82.1	77.5	84.6	72.4	69.7	75.7
Mehta <i>et al.</i> <sup>3+M+L+C</sup> [31]*	83.8	75.0	77.8	77.5	55.1	80.4	72.5	–	–	–	75.2
Luo <i>et al.</i> <sup>3+M+H</sup> [11], [77]	<b>95.5</b>	82.3	89.9	84.6	66.5	92.0	93.0	–	–	–	84.3
Nibali <i>et al.</i> <sup>3+M</sup> [12]	–	–	–	–	–	–	–	–	–	–	87.6
2.5D baseline <sup>3+M</sup>	95.1	90.7	86.8	<b>92.4</b>	<b>74.2</b>	94.1	91.7	92.1	89.0	<b>87.7</b>	89.9±0.2
<b>MeTRo (ours)</b> <sup>3+M</sup>	95.0	<b>91.8</b>	<b>90.2</b>	92.1	73.4	<b>95.1</b>	<b>91.8</b>	<b>93.4</b>	<b>90.3</b>	86.5	<b>90.6±0.4</b>
<i>Metric-scale skeletons (full scale recovery task)</i>											
2.5D baseline <sup>3+M</sup>	93.1	<b>89.3</b>	83.6	<b>93.1</b>	<b>73.7</b>	<b>93.2</b>	<b>91.1</b>	89.0	<b>87.9</b>	<b>89.4</b>	88.7±0.6
<b>MeTRo (ours)</b> <sup>3+M</sup>	<b>94.0</b>	89.2	<b>87.1</b>	89.1	68.9	92.6	90.3	<b>90.1</b>	87.8	85.7	88.2±0.5
<b>88.7±0.7</b>											

TABLE 4

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	<b>89.6</b>	52.1	<b>80.6</b>
<b>MeTRo (ours)</b>	<b>97.0</b>	<b>68.6</b>	<b>49.3</b>	<b>89.6</b>	<b>52.6</b>	81.1

TABLE 5

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	185.0	124.7	<b>77.8</b>
Present joints	338.0	332.5	173.6	76.8	<b>59.8</b>

ratio on MuCo is 1.22 and 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.

## 7.1 Inference Speed

Our method is capable of real-time inference. The root-relative architecture can process 511 crops per second on

TABLE 6  
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	4
MPJPE	normal strides	53.1	52.5	52.7	52.9
	center-aligned	50.9	50.2	50.0	<b>49.3</b>
Speed (crop per sec.)	no batching	160	150	105	38
	batch size 8	<b>511</b>	475	292	92

TABLE 7  
Augmentation ablation on H36M.

Geometry	Color	Occlusion	MPJPE
✓	–	–	58.0
✓	✓	–	52.8
✓	✓	✓	<b>49.3</b>

TABLE 8  
Results on MuPoTS-3D. Detected, unnormalized poses, no bone rescaling. (\*Re-evaluated public results; joint count: †17, ‡16, else 14)

	A-MPJPE↓	MPJPE↓	A-PCK↑	PCK↑	Det.Rate↑
Rogez <i>et al.</i> [74]	–	146 <sup>‡</sup>	–	–	86
Mehta <i>et al.</i> [31]	–	132 <sup>‡</sup>	–	–	93
Baseline in [39]	320 <sup>†</sup>	122 <sup>‡</sup>	–	–	91
Véges <i>et al.</i> [39]	292 <sup>†</sup>	120 <sup>‡</sup>	–	–	91
Véges <i>et al.</i> [75]*	257.2 (255 <sup>†</sup> )	119.4 (108 <sup>‡</sup> )	38.1	75.4	93
2.5D baseline	317.6 (313.6 <sup>†</sup> )	114.0 (110.0 <sup>‡</sup> )	<b>40.0±1.0</b>	79.3±0.3	<b>94.2±0.0</b>
MeTRAbs	<b>248.2 (246.9<sup>†</sup>)</b>	<b>108.2 (104.3<sup>‡</sup>)</b>	<b>40.2±1.9</b>	<b>81.1±0.4</b>	94.1±0.1
w/o abs. loss	328.8 (327.8 <sup>†</sup> )	108.4 (104.7 <sup>‡</sup> )	36.7±3.2	80.9±0.4	94.1±0.1

an RTX 2080 Ti desktop GPU when operating on batches of 8 crops at stride 32 (Tab. 6). Varying the heatmap resolution using dense prediction provides diminishing returns in accuracy (Tab. 6), 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

TABLE 9

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↑	PCK↑	A-PCK↑	PCK↑
F	F	37.2±1.7	76.2±0.5	39.3±1.7	79.9±0.5
F	W	<b>39.4</b> ±1.6	76.2±0.5	<b>41.6</b> ±1.6	80.0±0.5
W	F	35.6±1.8	77.1±0.4	37.6±1.8	81.0±0.5
W	W	38.1±1.8	<b>77.2</b> ±0.4	40.2±1.9	<b>81.1</b> ±0.4
-	F	33.0±3.3	77.0±0.4	34.9±3.3	80.8±0.4
-	W	34.8±3.1	77.0±0.4	36.7±3.2	80.9±0.4

TABLE 10

Results on the 3DPW challenge dataset. (PA=Procrustes analysis)

	MPJPE↓	MPJPE-PA↓	PCK@50mm↑	AUC@200mm↑
<i>Known association to GT identity</i>				
Sun <i>et al.</i> [78]	81.8	58.6	37.3	59.9
Kissos <i>et al.</i> [79]	83.2	59.7	42.4	62.3
MeTRAbs (ours)	<b>68.8</b>	<b>49.7</b>	<b>48.8</b>	<b>66.8</b>
<i>Unknown association to GT identity</i>				
MeTRAbs (ours)	85.1	56.7	45.8	63.2

per frame, respectively. The above calculations assume the image crops are available instantly and the time cost of detection is excluded.

## 7.2 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. Tab. 10 compares results using the 3DPW protocol. 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.

## 8 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.

## ACKNOWLEDGMENTS

We thank Yinglun Liu for help in evaluating on MuPoTS-3D. This work was funded, in parts, by a Bosch Research Foundation grant, the ERC Consolidator Grant project “DeeViSe” (ERC-CoG-2017-773161) and the EU H2020 projects ILIAD (H2020-ICT-2016b-732737) and CROWDBOT (H2020-ICT-2017-779942). Compute resources were granted by RWTH Aachen University under project “rwth0479”.

## REFERENCES

- [1] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, “3D human pose estimation in RGBD images for robotic task learning,” in *ICRA*, 2018.
- [2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, “Video based reconstruction of 3D people models,” in *CVPR*, 2018.
- [3] V. Srivastav, T. Issenhuber, A. Kadkhodamohammadi, M. de Mathelin, A. Gangi, and N. Padoy, “MVOR: A multi-view RGB-D operating room dataset for 2D and 3D human pose estimation,” in *MICCAI LABELS Workshop*, 2018.
- [4] N. Neverova, R. A. Güler, and I. Kokkinos, “Dense pose transfer,” in *ECCV*, 2018.
- [5] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in *ECCV*, 2016.
- [6] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, “Learning feature pyramids for human pose estimation,” in *CVPR*, 2017.
- [7] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, “Multi-scale structure-aware network for human pose estimation,” in *ECCV*, 2018.
- [8] J. Martinez, R. Hossain, J. Romero, and J. J. Little, “A simple yet effective baseline for 3D human pose estimation,” in *ICCV*, 2017.
- [9] D. Mehta *et al.*, “Vnect: Real-time 3D human pose estimation with a single RGB camera,” *ACM Trans. Graphics*, 2017.
- [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 *ICCV*, 2017.
- [11] C. Luo, X. Chu, and A. Yuille, “OriNet: A fully convolutional network for 3D human pose estimation,” in *BMVC*, 2018.
- [12] A. Nibali, Z. He, S. Morgan, and L. Prendergast, “3D human pose estimation with 2D marginal heatmaps,” in *WACV*, 2019.
- [13] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, “Coarse-to-fine volumetric prediction for single-image 3D human pose,” in *CVPR*, 2017.





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).

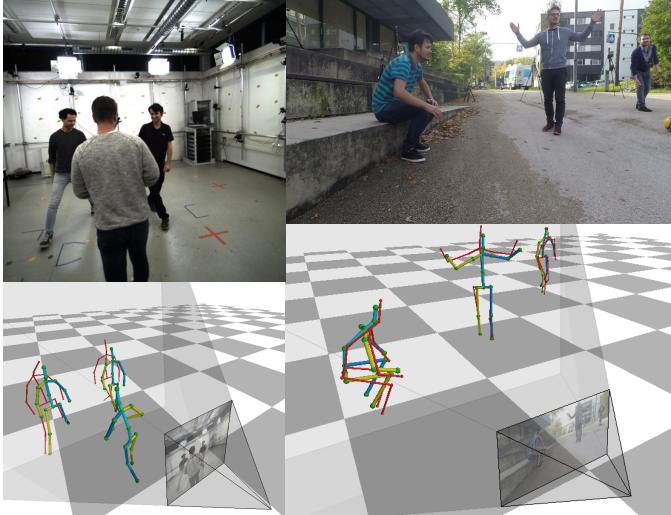


Fig. 8. Qualitative results on MuPoTS-3D (prediction in blue-yellow, ground truth in red).

and I will tell you your height, weight and body mass index," in *ICPR*, 2018.

- [34] 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," in *3DV*, 2017.
- [35] R. Dabral, N. B. Gundavarapu, R. Mitra, A. Sharma, G. Ramakrishnan, and A. Jain, "Multi-person 3D human pose estimation from monocular images," in *3DV*, 2019.
- [36] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgarheib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, and C. Theobalt, "XNect: Real-time multi-person 3D human pose estimation with a single RGB camera," *arXiv:1907.00837*, 2019.
- [37] G. Moon, J. Y. Chang, and K. M. Lee, "Camera distance-aware

top-down approach for 3D multi-person pose estimation from a single RGB image," in *ICCV*, 2019.

- [38] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, "Synthetic occlusion augmentation with volumetric heatmaps for the 2018 ECCV PoseTrack challenge on 3D human pose estimation," *arXiv:1809.04987*, 2018.
- [39] M. Véges and A. Lőrincz, "Absolute human pose estimation with depth prediction network," in *IJCNN*, 2019.
- [40] W. Jiang, N. Kolotouros, G. Pavlakos, X. Zhou, and K. Daniilidis, "Coherent reconstruction of multiple humans from a single image," in *CVPR*, 2020.
- [41] M. Fieraru, M. Zanfir, E. Oneata, A.-I. Popa, V. Olaru, and C. Sminchisescu, "Three-dimensional reconstruction of human interactions," in *CVPR*, 2020.
- [42] C. Ionescu, F. Li, and C. Sminchisescu, "Latent structured models for human pose estimation," in *ICCV*, 2011.
- [43] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *PAMI*, 2014.
- [44] A. Kar, S. Tulsiani, J. Carreira, and J. Malik, "Amodal completion and size constancy in natural scenes," in *ICCV*, 2015.
- [45] S. Park, S.-b. Lee, and J. Park, "Data augmentation method for improving the accuracy of human pose estimation with cropped images," *Pattern Recognition Letters*, 2020.
- [46] S. Vosoughi and M. A. Amer, "Deep 3D human pose estimation under partial body presence," in *ICIP*, 2018.
- [47] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *ECCV*, 2016.
- [48] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *PAMI*, 2018.
- [49] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014.
- [50] 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 *CVPR*, 2018.

- [51] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.
- [52] M. A. Islam, S. Jia, and N. D. Bruce, "How much position information do convolutional neural networks encode?" in *ICLR*, 2020.
- [53] S. Workman, C. Greenwell, M. Zhai, R. Baltenberger, and N. Jacobs, "Deepfocal: A method for direct focal length estimation," in *ICIP*, 2015.
- [54] Y. Liang, R. Ranade, S. Wang, D. Bai, J. Lee, M. Valtonen Ornhag, C. Olsson, A. Heyden, Z. Gao, J. Zhang *et al.*, "The "vertigo effect" on your smartphone: Dolly zoom via single shot view synthesis," in *CVPR Workshops*, 2020.
- [55] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose: Supplementary material," in *CVPR*, 2017.
- [56] B. Amos and J. Z. Kolter, "OptNet: Differentiable optimization as a layer in neural networks," in *ICML*, 2017.
- [57] A. Agrawal, B. Amos, S. Barratt, S. Boyd, S. Diamond, and J. Z. Kolter, "Differentiable convex optimization layers," in *NeurIPS*, 2019.
- [58] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv:1804.02767*, 2018.
- [59] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *ICCV*, 2017.
- [60] D. Luvizon, D. Picard, and H. Tabia, "Multi-task deep learning for real-time 3D human pose estimation and action recognition," *PAMI*, 2020.
- [61] H.-S. Fang\*, Y. Xu\*, W. Wang, X. Liu, and S.-C. Zhu, "Learning pose grammar to encode human body configuration for 3D pose estimation," in *AAAI*, 2018.
- [62] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, "3D human pose estimation in the wild by adversarial learning," in *CVPR*, 2018.
- [63] G. Pavlakos, X. Zhou, and K. Daniilidis, "Ordinal depth supervision for 3D human pose estimation," in *CVPR*, 2018.
- [64] D. Liu, Z. Zhao, X. Wang, Y. Hu, L. Zhang, and T. Huang, "Improving 3D human pose estimation via 3D part affinity fields," in *WACV*, 2019.
- [65] J. Xu, Z. Yu, B. Ni, J. Yang, X. Yang, and W. Zhang, "Deep kinematics analysis for monocular 3D human pose estimation," in *CVPR*, 2020.
- [66] S. Sharma, P. T. Varigonda, P. Bindal, A. Sharma, and A. Jain, "Monocular 3D human pose estimation by generation and ordinal ranking," in *ICCV*, 2019.
- [67] Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, and N. M. Thalmann, "Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks," in *ICCV*, 2019.
- [68] B. X. Nie, P. Wei, and S. Zhu, "Monocular 3D human pose estimation by predicting depth on joints," in *ICCV*, 2017.
- [69] I. Habibie, W. Xu, D. Mehta, G. Pons-Moll, and C. Theobalt, "In the wild human pose estimation using explicit 2D features and intermediate 3D representations," in *CVPR*, 2019.
- [70] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *AAAI*, 2020.
- [71] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," <http://host.robots.ox.ac.uk/pascal/VOC/voc2012/>, 2012.
- [72] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, "How robust is 3D human pose estimation to occlusion?" in *IROS Workshop - Robotic Co-workers 4.0*, 2018.
- [73] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *ECCV*, 2008.
- [74] G. Rogez, P. Weinzaepfel, and C. Schmid, "LCR-Net: Localization-classification-regression for human pose," in *CVPR*, 2017.
- [75] M. Veges and A. Lorincz, "Multi-person absolute 3D human pose estimation with weak depth supervision," *arXiv:2004.03989*, 2020.
- [76] K. Zhou, X. Han, N. Jiang, K. Jia, and J. Lu, "HEMlets pose: Learning part-centric heatmap triplets for accurate 3D human pose estimation," in *ICCV*, 2019.
- [77] C. Luo, X. Chu, and A. Yuille, "OriNet-demo," <https://github.com/chenxuluo/OriNet-demo>, 2018, [accessed 16-Nov-2018].
- [78] Y. Sun, Q. Bao, W. Liu, Y. Fu, and T. Mei, "CenterHMR: a bottom-up single-shot method for multi-person 3D mesh recovery from a single image," *arXiv:2008.12272*, 2020.
- [79] I. Kissos, L. Fritz, M. Goldman, O. Meir, E. Oks, and M. Kliger, "Beyond weak perspective for monocular 3D human pose estimation," *arXiv:2009.06549*, 2020.
- [80] A. Benzine, B. Luvison, Q. C. Pham, and C. Achard, "Single-shot 3D multi-person pose estimation in complex images," *Pattern Recognition*, 2020.
- [81] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *CVPR*, 2017.
- [82] Y.-T. Hu, H.-S. Chen, K. Hui, J.-B. Huang, and A. G. Schwing, "SAIL-VOS: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines," in *CVPR*, 2019.
- [83] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews *et al.*, "Panoptic studio: A massively multiview system for social interaction capture," *TPAMI*, 2017.



**István Sárándi** is a PhD candidate in the Computer Vision group of RWTH Aachen University, supervised by Prof. Dr. Bastian Leibe. He holds a BSc degree in Computer Engineering from the Budapest University of Technology and Economics, and an MSc degree in Computer Science from RWTH Aachen University. 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** defended his PhD thesis on multimodal human detection, tracking and analysis for robots in crowded environments at the University of Freiburg, Germany in 2020. Since 2016, he is a research scientist in autonomous systems and robot perception at Bosch Corporate Research. 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 has co-authored peer-reviewed publications at major international conferences and journals, served on different program committees in robotics and AI, and received an outstanding reviewer award at ICRA 2019.



**Kai Oliver Arras** is the head of robotics research and chief expert in robotics at Robert Bosch GmbH. Until 2015, he was assistant professor for social robotics and HRI at the University of Freiburg where he was awarded a DFG Junior Research Group Leader Grant. He obtained his PhD degree from EPFL and was a post-doctoral researcher at KTH Stockholm and at the University of Freiburg. He published around 120 peer-reviewed papers, articles, editorials and book chapters on robot navigation, perception, planning, system integration and was member of various program committees in robotics, AI, HRI and computer vision.



**Bastian Leibe** is a full Professor of Computer Science at RWTH Aachen University, Germany, where he leads the Computer Vision group. He holds an MS degree from the Georgia Institute of Technology (1999), a Diploma degree from the University of Stuttgart (2001), and a PhD from ETH Zurich (2004), all three in Computer Science. His main research interests are in computer vision and machine learning for dynamic visual scene understanding, encompassing object recognition, tracking, segmentation, and 3D reconstruction. He has published over 130 articles in peer-reviewed journals and conferences. 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. In 2012, he was awarded a European Research Council (ERC) Starting Grant, and in 2017 an ERC Consolidator Grant. He has been Program Chair for ECCV 2016 and Area Chair and program committee member for all major computer vision conferences.