Absolute-ROMP: Recovering Multi-person 3D Poses and Shapes with Absolute Scales from a Single RGB Image

Bilal AbdulRahman[1(✉)] and Zhigang Zhu[2]

[1] The Graduate Center, The City University of New York, New York, NY 10016, USA
babdulrahman@gradcenter.cuny.edu

[2] The City College and The Graduate Center, The City University of New York, New York, NY 10031, USA

Abstract. One of the grand challenges in computer vision is to recover 3D poses and shapes of multiple human bodies with absolute scales from a single RGB image. The challenge stems from the inherent depth and scale ambiguity from a single view. The state of the art on 3D human pose and shape estimation mainly focuses on estimating the 3D joint locations relative to the root joint, defined as the pelvis joint. In this paper, a novel approach called Absolute-ROMP is proposed, which builds upon a one-stage multi-person 3D mesh predictor network, ROMP, to estimate multi-person 3D poses and shapes, but with absolute scales from a single RGB image. To achieve this, we introduce absolute root joint localization in the camera coordinate frame, which enables the estimation of 3D mesh coordinates of all persons in the image and their root joint locations normalized by the focal point. Moreover, a CNN and transformer hybrid network, called TransFocal, is proposed to predict the focal length of the image's camera. This enables Absolute-ROMP to obtain absolute depth information of all joints in the camera coordinate frame, further improving the accuracy of our proposed method. The Absolute-ROMP is evaluated on the root joint localization and root-relative 3D pose estimation tasks on publicly available multi-person 3D pose datasets, and TransFocal is evaluated on a dataset created from the Pano360 dataset. Our proposed approach achieves state-of-the-art results on these tasks, outperforming existing methods or has competitive performance. Due to its real-time performance, our method is applicable to in-the-wild images and videos.

Keywords: Machine learning · Computer vision · 3D reconstruction · Camera calibration · Pose prediction · Human mesh regression

1 Introduction

Three-Dimensional (3D) human pose and shape estimation is one of the most active research topics within the current landscape of computer vision and machine learning,

The work is supported by AFOSR Dynamic Data Driven Applications Systems (Award #FA9550-21-1-0082). The work is also supported in part by NSF via the Partnerships for Innovation Program (Award #1827505) and the CISE-MSI Program (Award #2131186), and ODNI via the Intelligence Community Center for Academic Excellence (IC CAE) at Rutgers University (Awards #HHM402-19-1-0003 and #HHM402-18-1-0007).

A. A. de Sousa et al. (Eds.): VISIGRAPP 2023, CCIS 2103, pp. 73–97, 2024.
https://doi.org/10.1007/978-3-031-66743-5_4

thanks to its many applications in various fields. These include robotics [12,64], activity recognition [7,44], graphics [2,6] and human-object interaction detection [13,36,37,52]. Current approaches on 3D human pose and shape estimation tend to mainly focus on the estimation of the 3D joint locations relative to the root joint, usually defined as the one closest to the shape centroid. In case of humans, it is defined as the pelvis joint.

This paper aims to address the problem of estimating the absolute 3D poses and shapes of multiple people simultaneously from a single RGB image. It is quite a challenge to accurately recover 3D poses and shapes of multiple persons with absolute scales from a single RGB image, due to the inherent depth and scale ambiguity of a single view. In order to addressing this ambiguity, assessing various spatial cues in the image as a whole is required, such as scene layouts, body dimensions, and inter-person relationships. Compared to existing approaches to the 3D pose and shape estimation problem that focuses on recovering the root-relative pose, the task addressed here additionally needs to recover the 3D translation of each person in the camera coordinate system (or sometimes called camera coordinate frame).

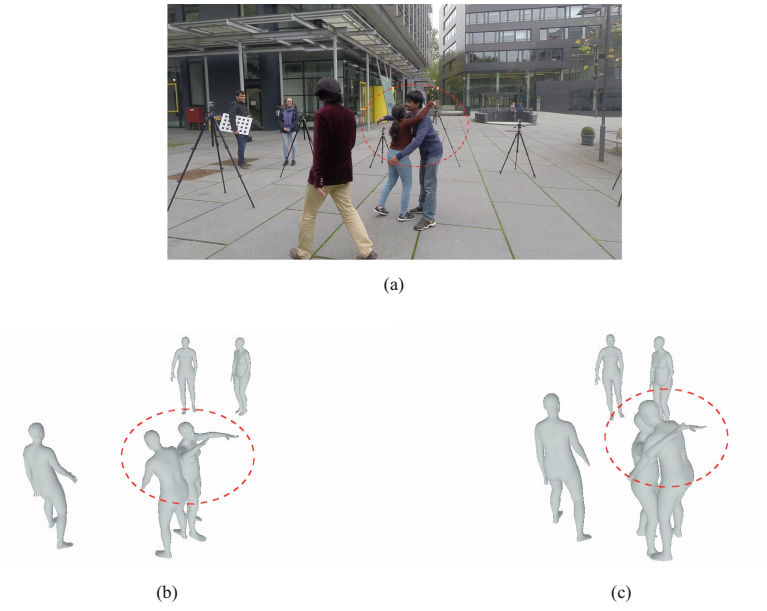

Fig. 1. Absolute-ROMP is able to correctly position two people hugging: (a). Original image. (b). ROMP mesh positioning using camera parameters as in [55]. (c). Absolute-ROMP mesh positioning using absolute depth prediction. This figure is modified from a figure in the paper we presented at VISAPP 2023 [1].

To further motivate the work, we would like to point out that estimating the absolute 3D location of each person in an image is essential for understanding human-to-human interactions (e.g., two people hugging in Fig. 1). However, the task is also very challenging to estimate the absolute positions of multiple individuals, since multi-person

activities often take place in cluttered scenes, thus leading to inherent depth ambiguity and occlusions. Here we also argue that body dimensions alone only paint a vague picture of absolute depth. Robust estimation of global positions requires multiple information cues over the entire image, such as geometric cues, human body sizes in the image, any occlusions which may affect the perceived sizes of persons and the layout of the entire scene.

In the literature, most existing methods for absolute multi-person 3D pose estimation extend a single-person approach with an added step to recover the absolute position of each detected person individually. They either use another neural network to regress the 3D translation of a person from the cropped image [50] or compute it based on the prior knowledge about the body size [9], which ignores the global context of the whole image. While others employ complicated architecture with extensive steps, drastically slowing down inference [40].

For an efficient one-stage estimation, we build upon ROMP [55], a light weight, accurate end to end multi-person 3D mesh prediction network, by adding in absolute root joint depth estimation and localization head, while maintaining its end to end and light weight nature. We thus call our network *Absolute-ROMP*. Even though we adopt the same end to end pipeline of ROMP for the task of multi-person absolute 3D mesh estimation, by leveraging depth cues from the entire scene and prior knowledge of the typical size of the human pose and body joints, we can estimate the depth of a person in a monocular image with considerably high accuracy. The target depths are discretized into a preset number of bins, in order to limit the range of predictions and thus improve the prediction performance. The range of these bins is chosen after taking prediction error mitigation and reasonable distance estimation in consideration. We employ a soft-argmax operation on the bins for improved accuracy as compared to exact bin locations, and for faster convergence during training without losing precision to direct numerical regression. We also perform experiments on different bin sizes and even compare with numerical output and choose the best performing method.

There is one key issue with absolute location that causes most other works to avoid it: it requires the knowledge of intrinsic camera parameters for accurate prediction. Since different focal lengths lead to different sizes of the same person in the image. Therefore, we also design and train a hybrid network called *TransFocal*, which integrates a CNN and a vision transformer to predict the vertical field of view of the image and thus the focal length. With the predicted focal length, we can estimate absolute distance in camera coordinates without the need for known camera intrinsic parameters. Our Transfocal model uses embeddings from a CNN model ResNet [16], which are then converted to tokens to be fed into the vision transformer [10]. The added perceptual grouping and self-attention from the transformer gives our network an edge in its accuracy over previous work [30]. As shown in Fig. 1, Absolute-ROMP achieves better accuracy than ROMP in positioning the poses.

In summary, our contributions in this work are:

– A revised network Absolute-ROMP, with an absolute depth estimation head is proposed for the ROMP network using a combination loss. It takes input from the backbone and then predicts the absolute location of each person in the frame normalized by the focal length. To achieve this, we introduce absolute root joint localization in

the camera coordinate frame, which enables the estimation of 3D mesh coordinates of all persons in the image and their root joint locations normalized by the focal length. This single-shot approach allows the system to better learn and reason about the inter-person depth relationship, leading to improved multi-person 3D estimation.
– A novel network TransFocal is designed and trained to predict focal length of the image, thus negating the requirement of intrinsic parameters. This focal length is then multiplied with the output from the depth head to get the final output. This enables us to obtain absolute depth information of all joints in the camera coordinate frame, further adding to our proposed method.
– Quantitative and qualitative results show that our approach outperforms or has competitive performance to the state-of-the-art approaches on multiple benchmark datasets, under various evaluation metrics. We evaluate Absolute-ROMP on the root joint localization and root-relative 3D pose estimation tasks on publicly available multi-person 3D pose datasets. We also evaluate TransFocal on a dataset created from the Pano360 dataset. Our proposed approach achieves state-of-the-art results on these tasks, outperforming existing methods. Additionally, our method is applicable to in-the-wild images and videos due to its real-time performance.

This is an extended version of the paper we presented at VISAPP 2023 [1]. In this extended version, we reorganize the sections of the paper (especially the Methodology section) and provide more detailed explanations of the new contributions of our proposed method. We also introduce new experiments in the ablation study of binning versus numerical output as well as various numbers of bins in TransFocal, and provide a through analysis of the inference time of the entire system and individual components to justify the effectiveness and efficiency of our approach. Any table, figure or equation used in the conference paper is referenced with a citation in the caption.

This paper is organizes as the follows. Section 2 discusses related work. Section 3 describes the proposed Absolute-ROMP, including the absolute depth map head and the focal length estimation network: TransFocal. Section 4 provides some key implementation details in network architectures and training/testing settings. Section 5 presents experimental results and ablation studies. Section 6 provides a few concluding remarks.

2 Related Work

2.1 Single-Person 3D Mesh Regression

In single-person 3D mesh regression, parametric human body models have been widely adopted since they allow regression of the 3D meshes from images. A good example of such models is the Skinned Multi-Person Linear Model (SMPL) [45]. The key is that these models allow complex 3D human mesh vertices to be encoded into low dimensional parameter vectors. Various weakly supervised approaches have been used, which lead to reasonable accuracy in single-person 3D mesh regression, using various cues, such as semantic segmentation [59], geometric prior [23], motion analysis [24,27] and 2D human pose [8]. A part-guided attention mechanism is used in [28] in order to overcome occlusions. This is done by exploiting information about the visibility of individual body parts while leveraging information from neighboring body-parts to predict

occluded parts. In [30], predicted camera calibration parameters are used to aid in the regression of the body mesh parameters.

2.2 Multi-person 3D Pose and 3D Mesh Estimation

3D pose refers to a person's joint positions in 3 dimensional space where as 3d mesh also takes into account the shape of the person recreating a 3D mesh of the person's body using a body model such as SMPL [45]. Extending the work from a single person to multiple persons is more challenging but more useful. For multi-person 3D pose estimation, various approaches have been proposed. In [49], occlusion-robust pose-maps are proposed to exploit the body part association to avoid bounding box prediction. In [5], an anchor-based one-stage model is used, which relies on a huge number of pre-defined anchor predictions and positive anchor selections. To handle person-to-person occlusions, a single-shot system SMAP is proposed in [62], which first regresses a set of 2.5D representations of body parts and then reconstructs the 3D absolute poses based on these 2.5D representations with a depth-aware part association algorithm. Top-down designs are employed in both [54] and [50], which estimate targets via regression from anchor-based feature proposals.

For further multi-person 3D mesh estimation, most approaches follow a multi-stage design. Built on Faster-RCNN [53], a network called Coherent Reconstruction of Multiple Humans (CRMH) is proposed in [19]. The RoI-aligned feature of each person is used to predict the SMPL parameters as in [45]. In [61], the 3D mesh of each person is estimated from its intermediate 3D pose estimation. Their work further employs multiple scene constraints to optimize the multi-person 3D mesh results. In all these methods, the complex multi-step process requires a repeated feature extraction, which is computationally expensive.

The ROMP network proposed in [55] regresses meshes in a one-stage fashion for multiple 3D people (thus termed ROMP). ROMP [55] learns an explicit one-stage pixel-level representation with a holistic view, which improves both the accuracy and efficiency in multi-person in-the-wild scenes. Therefore, our proposed model is based on ROMP.

2.3 Monocular Absolute Depth Estimation

Estimating depth information from a single view suffers from inherent ambiguity. Nevertheless, several methods make remarkable advances in the last few years [32,39]. In [38], a dataset for depth estimation is obtained by employing the frozen poses and the moving camera of the "mannequin challenge". Training data is generated by using multi-view stereo reconstruction and a data-driven approach is adopted to recover a dense depth map. However, the depth maps generated lack scale consistency and therefore do not reflect the real depths. As described above, the SMAP approach in [62] first regresses a set of 2.5D representations of body parts and then reconstructs their 3D absolute poses with a depth-aware part association algorithm. In [40], the 2D pose of a person is estimated with heatmaps of the joints, which are used as attention masks for pooling features from image regions corresponding to the target person. To predict the depth of each joint, a skeleton-based Graph Neural Network (GNN) is used. With

a coarse-to-fine architecture, an integrated model is used in [34] to estimate human bounding boxes, human depths, and root-relative 3D poses simultaneously.

All these methods either employ multi-step prediction or use large networks, which slow down the inference. Our method is able to estimate absolute depths and regress 3D meshes of multiple people in real time with high accuracy, while also regressing the shape parameters of individual persons.

2.4 Focal Length and Other Camera Parameter Estimation

Estimating camera parameters from a single image became popular recently [17,25, 57,58,63]. In order to estimate camera rotations and fields of view (FOVs), these methods train neural networks for leveraging geometric cues in the image. Using an AlexNet backbone, the approach proposed in [57] regresses the horizontal field of view. Apart from [57], other methods discretize the continuous space of rotations into bins, casting the problem as a classification task, and applying cross entropy [58] or KL-divergence [17,63] losses. Also using a binning technique, [30] trains a neural network with a bespoke-biased loss on a new collected dataset.

None of these methods take advantage of the latest architecture innovation in the vision space, i.e., transformer networks [10]. [33] is a rare example that takes both an image and line segments as input and regresses the camera parameters based on the transformer encode-decoder architecture. The line segments are extracted from the input image using the LSD algorithm [15], and then mapped to geometric tokens which are generated by a transformer encoder. The subsequent transformer decoder aggregates both semantic and geometric tokens along with the queries for the camera parameters. In contrast, our model *only* employs vision transformer to encode the features from a single image and then a simple MLP layer is used for decoding, thus leading to an integration of a vision transformer and a CNN hybrid network with a combination of losses during supervision.

3 Methodology

Since ourt Absolute-ROMP is an extension of the ROMP network [55], we will explain the working of ROMP before going into details about our addition. Figure 2 shows the overall system diagram. ROMP regresses meshes in a one-stage fashion for multiple 3D people. In the same way, Absolute-ROMP employs a one-stage, multi-head design with a HRNeT-32 backbone [56] followed by 4 head networks: Body Center Map, Camera Map, SMPL Map and our newly-designed Root Depth Map. We also maintain Coord-Conv [43] from ROMP to enhance the spatial information. Therefore, the backbone feature is the combination of a coordinate index map and output feature embeddings from HRNET-32. Details of the backbone model will be described in Sect. 4.

Given a RGB image as input, the backbone HRNet-32 generates a feature set that is used as input to the four heads for complete end to end prediction: a body center heatmap, camera parameters, SMPL parameters and a root depth map, all in the camera coordinate frame. A separate focal length estimation network - TransFocal - is designed (and trained separately) to estimate the focal length of the image in order to

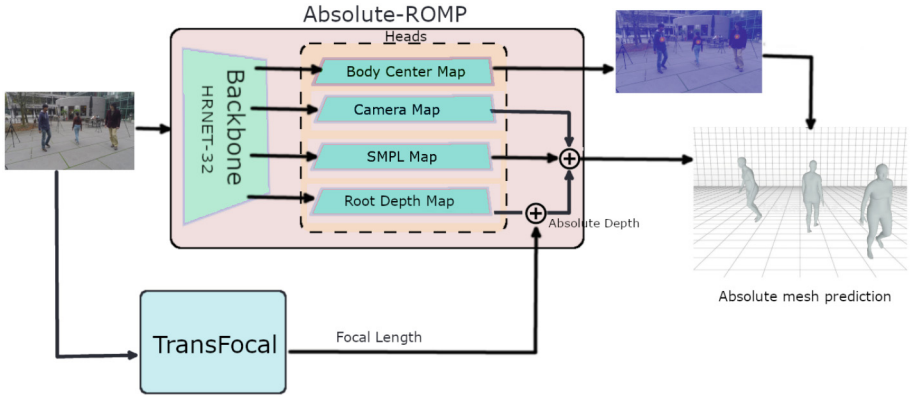

Fig. 2. Overview of how the system works [1]. Absolute-ROMP predicts the mesh parameters and depth. The focal length is predicted by TransFocal which are then used to get the complete absolute 3D coordinates. This figure is modified from a figure in the paper we presented at VISAPP 2023 [1], by adding "Absolute Depth" output as an integration of results of Root Depth Map and Tansfocal on the figure.

de-normalize the output from the Root Depth Map, and to generate an absolute root depth map. This absolute root depth map along with the SMPL parameters from the SMPL Map and the camera parameters from the Camera Map, enables us to create absolute 3D body meshes which are then correctly filtered with the help of the Body Center Map to generate the final output. In our experiments, the resolution for each map is 64×64.

This section will start with the description of a concise body-center guided representation introduced in [55], then describe the four head networks with an emphasis on our extension (i.e., the absolute depth map head) leading to the Absolute-ROMP, and finally discuss our new technical contributions in parameter sampling (with improved relative depth estimation), loss function design (with the new root depth loss) and the proposed focal length estimation - TransFocal. Figure 3 illustrates several key steps of the Absolute-ROMP: the body center headmap in (a), the SMPL parameters and the root depth prediction in (b), and the final result in (c).

3.1 Preliminary: Collision Aware Representation

The entire Absolute-ROMP framework is built upon a concise body-center guided representation. Defining an explicit and robust body center is crucial to enable the model to accurately estimate the center location in various scenarios. Utilizing the bounding box center of a person as the body center holds little relevance due to its propensity to lie outside the bodily region and its lack of alignment with any precise anatomical point. To ensure consistent parameter sampling, it becomes imperative to establish an explicit body center. Henceforth, we computationally derive each body center by leveraging the ground truth 2D pose of a person.

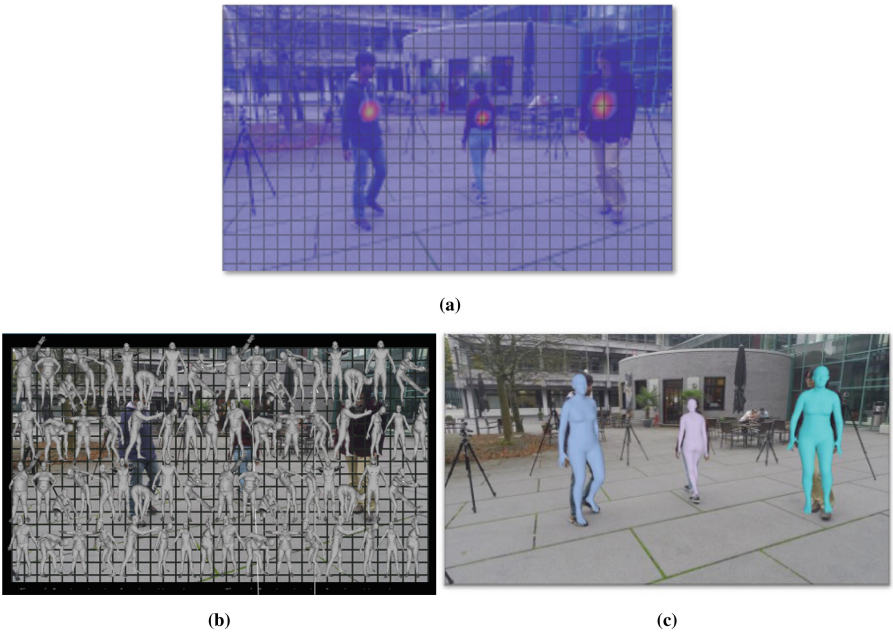

Fig. 3. (a). An illustration of how the body center heatmap encodes size information into the Gaussian representation (b). Shows how SMPL parameters and root depth are predicted across the grid (c)Final result after parameter sampling.

Given that occlusion of body joints is a common occurrence, our approach involves defining the body center as the central point among the observable torso joints, encompassing the neck, left and right shoulders, pelvis, and left and right hips. In instances where all torso joints remain hidden, we determine the center by calculating the average position of the visible joints. This method encourages the model to estimate the body's location based on the discernible parts, even when certain joints are obscured.

Nonetheless, challenges arise in scenarios involving densely packed individuals, where severe overlap may lead to close proximity or even coinciding body centers within the image plane. The resulting collision problem introduces ambiguity and poses difficulties in discerning individual centers in crowded situations. To tackle this predicament, ROMP [55] introduces a more resilient representation known as the Collision Aware Representation (CAR). To mitigate the predicament arising from intertwined body centers, ROMP incorporates a repulsion field within the CAR framework. Within this field, each body center assumes the role of a positive charge, characterized by a repulsion radius equivalent to its Gaussian kernel size (for details, please see Eq. 4 in Sect. 3.2 below). The intensity of repulsion between two adjacent body centers increases as their proximity intensifies, thereby compelling them to distance themselves from each other.

Let us consider c_1 and c_2 as the body centers of two overlapping individuals. If the Euclidean distance between them, denoted as d_{cm}, satisfies the condition $d_{cm} < k_1 + k_2 + 1$, the repulsion mechanism is triggered, prompting the close centers to separate. This is achieved through the following equations:

$$\hat{c}_1 = c_1 + \gamma d_p, \tag{1}$$

$$\hat{c}_2 = c_2 - \gamma d_p, \tag{2}$$

$$d_p = \frac{k_1 + k_2 + 1 - d_{cm}}{d_{cm}}(c_1 - c_2), \tag{3}$$

Here, d_p denotes the repulsion vector originating from c_2 and directed towards c_1, and γ signifies an intensity coefficient that modulates the strength of repulsion. When multiple individuals overlap, we apply the same equations to compute repulsion vectors d_i^p for each pair of centers. For a center affected by N repulsive forces, we calculate the resultant composition of these forces by summing them numerically, expressed as $\sum_{i=1}^{N} d_i^p$. During the training phase, CAR is employed to disentangle closely positioned body centers, facilitating more accurate localization.

3.2 The Four Head Networks

Body Center Map Head. To generate the Body Center heatmap, which represents the 2D human body center of each person in an image, we employ the method stated in [55] that incorporates scale information and increases the level of detail in the representation.

In ROMP's enhanced approach, each body center is represented as a Gaussian distribution within the heatmap. To facilitate better representation learning, we integrate the scale information of the body in the 2D image into the Body Center heatmap. This integration allows us to adapt the spread of the Gaussian distribution to capture individuals of different sizes more effectively. The Gaussian kernel size, denoted as k, is used to determine the spread or influence of the body center in the heatmap. We calculate k for each person's center based on their 2D body scale in the image. The kernel size is derived as follows:

$$k = k_l + \left(\frac{d_{bb}}{\sqrt{2}W}\right)^2 \cdot k_r \tag{4}$$

In Eq. 4, k_l represents the minimum kernel size, which establishes the baseline spread of the Gaussian distribution. The term $\left(\frac{d_{bb}}{\sqrt{2}W}\right)^2$ accounts for the ratio of the person's bounding box diagonal length, d_{bb}, to the width of the Body Center heatmap, W. Squaring this ratio ensures a proportional increase in the spread of the Gaussian distribution. Additionally, k_r serves as a variation factor, allowing fine-tuning of the spread based on specific characteristics or requirements.

By incorporating the body scale information through the calculation of the Gaussian kernel size, the resulting Body Center heatmap provides a more detailed and nuanced representation of the human body centers in the 2D image. Figure 3(a) illustrates the integration of the scale information of three bodies into their Gaussian distributions. This enhancement facilitates more effective learning and analysis of body center-related features and patterns.

Camera Map Head. The Camera Map, denoted as W, encompasses the 3-dimensional camera parameters (s, tx, ty), which describe the 2D scale s and translation $t = (tx, ty)$ of each person in the image. The scale s reflects the body size and, to some extent, the depth. The translation parameters tx and ty range between -1 and 1 and represent the normalized translation of the human body relative to the image center along the x and y axes, respectively.

The 2D projection, denoted as $J_b = \{(x_{bk}, y_{bk})\}$, of the 3D body joints $J = \{(x_k, y_k)\}$ can be derived using the following equations:

$$x_{bk} = s \cdot x_k + tx; y_{bk} = s \cdot y_k + ty \tag{5}$$

In the above equations, x_k and y_k represent the coordinates of the k-th 3D body joint (k represent body joints which range from 1 to 22 the maximum), while x_{bk} and y_{bk} represent their respective projections in the 2D image. The scale factor s is applied to the coordinates to account for the body size, and the translation parameters tx and ty enable more accurate position estimates of the body joints relative to the image center. The utilization of translation parameters in the camera map allows for improved precision in position estimation compared to relying solely on the Body Center heatmap.

SMPL Map Head. The SMPL Map contains the 142-dimensional SMPL parameters of a body mesh, which describe the 3D pose and shape of the body mesh. The SMPL model establishes an efficient mapping from the pose θ and shape β parameters to the human 3D body mesh $M \in \mathbb{R}^{6890\times3}$. The *shape parameter* $\beta \in \mathbb{R}^{10}$ represents the top-10 PCA coefficients of the SMPL statistical shape space. The *pose parameters* $\theta \in \mathbb{R}^{6\times22}$ encompass the 3D rotations of the 22 body joints, represented in a 6D representation. In total the SMPL map has 142 dimensions $(10 + 6 \times 22)$.

In ROMP's implementation [55], a modified version of the SMPL model (where the last two hand joints are excluded) is employed. We also employ the same modified model. The 3D rotation of the first joint denotes the body's 3D orientation in the camera coordinate system, while the remaining rotations represent the relative 3D orientations of each body part with respect to its parent in a kinematic chain.

To derive the 3D joints J, we employ a linear mapping via the pose matrix The utilization of the SMPL Map allows us to represent the complex 3D pose and shape of the human body in a compact and efficient manner. By leveraging the pose and shape parameters, we can generate the corresponding 3D body mesh and derive the 3D joint locations.

Root Depth Map Head. Assuming that each location on the Root Depth Map represents the center of a human body, we aim to estimate the absolute depth of the corresponding root joint. Instead of directly regressing the numerical depth value, we employ a *binning technique* within the log depth space. The binning resolution is set to 120, chosen after considering prediction error mitigation and reasonable distance estimation based on all available data. There are two important treatments in the Root Depth Map head using the binning technique.

First, since different focal lengths of the camera can affect the scales of a person in the image, it becomes impractical to estimate absolute depth from images captured

by arbitrary cameras. If we want to simultaneously estimate focal lengths using the Root Depth Map Head, we have to use 3D human datasets that provide absolute depth information. However such datasets tend to have minimal variation in focal lengths, typically employing the same camera for all images within a given dataset. This presents a challenge for an integrated model to learn the variable focal lengths, as it can potentially overfit on the focal lengths present in the training datasets. In 3D pose and mesh estimation, the sizes of individuals may appear differently in images taken with cameras having different focal lengths. This would hinder prediction performance as model is not trained to generalize for focal length variations. To address these issues, we normalize (divide) the ground truth depth of an image by the ground truth focal length and therefore the Root Depth Map head is free from learning and predicting the focal length. The estimation of focal length is achieved through the training of a network called TransFocal, which we will elaborate on in a subsequent subsection. The absolute depth map is the final output after integrating the results from the Root Depth Map head and the estimated focal length from the TransFocal network in Fig. 2, in that the root depth map generated by the Root Depth Map head is normalized by the focal length.

Second, to enhance accuracy beyond exact integer bin values, we employ a soft-argmax operation on the bins. Exact bins provide integer outputs, limiting precision, while soft bins can output any number between bin indices based on the output value between 0 and 1 from the bin. Consequently, there is minimal precision loss, enabling the actual prediction to be a ratio of the bins, thereby improving the granularity and accuracy of the model. The computation of the bin index within the log depth space is as follows:

$$b(\hat{d}) = \frac{\log \hat{d} - \log S}{\log E - \log S}(N-1) \tag{6}$$

Here, $b(\hat{d})$ represents the bin index of the normalized depth $\hat{d}$. N denotes the total number of bins, and $[S, E]$ represents the range of the bins. If the output map size is 64×64, we can conceptualize this style of binning the depth map as a collection of 1D heatmaps, yielding 64×64 predictions for each image. In other words, we have a 64×64 collection of 1D heatmaps. The predicted bin values B are subsequently converted back to normalized depth using the following equation:

$$\hat{d} = \exp\left[\frac{\sum_{i=0}^{N-1} B_i \times i}{N-1}(\log E - \log S) + \log S\right] \tag{7}$$

Figure 3(b) shows how SMPL parameters and normalized absolute root depth are predicted across the grid.

3.3 Key Technical Issues and Solutions

In the following two subsections, we will detail the Absoulte-ROMP's implementation, highlighting our new improvements: In parameter sampling we emphasize our improved relative depth estimation approach. In loss function design, we introduce the root depth loss. Finally, in focal length estimation, we detail our bespoke TransFocal: a hybrid network with CNN and Transformer to estimate the focal length.

Parameter Sampling. To derive the 2D coordinates of a set of body centers $\{c\}$ from the estimated Body Center Map Cm and acquire the corresponding 3D body meshes and root depth, we employ a sequence of procedures encompassing center parsing, matching, and sampling. Initially, Cm acts as a probability map, where the presence of local maxima indicates potential body centers. To identify these local maxima, we apply the max pooling operation denoted as $M_p(Cm)$ and subsequently perform a logical conjunction with Cm, denoted as $Mp(Cm) \wedge Cm$. Consequently, 2D coordinates $\{c\}$ possessing confidence scores surpassing a designated threshold t_c are identified as local maxima. These confidence scores at each c are arranged in a ranked manner, with the top N centers being selected as the definitive set.

During the inference stage, we directly extract the SMPL parameters from the SMPL Map Pm at the corresponding identified centers c. However, during the training process, the estimated values of c are matched with the nearest ground truth body centers by employing the L2 distance as a measure of proximity.

Moreover, to approximate the relative depth order among multiple individuals, we initially leverage the center confidence inferred from Cm and the 2D body scale denoted as s, obtained from the camera parameters within the Camera Map Am. In scenarios where individuals exhibit discernible variations in scales, the individual with the larger s value is presumed to occupy the foreground position. Conversely, when individuals possess comparable scales, the person with a higher center confidence is considered to be located in the foremost position. After a certain loss threshold on the depth prediction, the absolute depth can be used as reliable measure for depth order and replaces the method stated above.

Figure 3(c) shows the final result of an image with the estimation of poses and shapes of three persons.

Loss Functions. To provide supervision for Absolute-ROMP, individual loss functions are employed for different maps. The supervision of Absolute-ROMP entails the utilization of the weighted sum of the body center loss L_c, mesh parameter loss L_p, and the root depth loss L_d.

Body Center Loss. The body center loss L_c promotes a high confidence value at the body center c within the Body Center heatmap Cm and a low confidence elsewhere. To address the imbalance between center and non-center locations in Cm, the Body Center heatmap is trained using the focal loss [41]. Given the predicted Body Center heatmap C_m^p and the ground truth C_m^{gt}, L_c is defined as follows:

$$L_c = -\frac{L_{pos} + L_{neg}}{\sum I_{pos}} w_c, \tag{8}$$

Here, $L_{neg} = \log(1 - C_m^p)(C_m^p)^2(1 - C_m^{gt})^4(1 - I_{pos})$ and $L_{pos} = \log(C_m^p)(1 - C_m^p)^2 I_{pos}$ represent the positive and negative focal loss terms, respectively. I_{pos} is a binary matrix with a positive value at the body center location, and w_c denotes the loss weight.

Mesh Parameter Loss: L_p is defined as follows:

$$L_p = w_{pose}L_{pose} + w_{shape}L_{shape} + w_{j3d}L_{j3d} + w_{paj3d}L_{paj3d} + w_{pj2d}L_{pj2d} + w_{prior}L_{prior}, \tag{9}$$

where L_{pose} represents the L2 loss of the pose parameters in the 3×3 rotation matrix format, L_{shape} represents the L2 loss of the shape parameters, L_{j3d} represents the L2 loss of the 3D joints J regressed from the body mesh M, L_{paj3d} represents the L2 loss of the 3D joints J after Procrustes alignment, L_{pj2d} represents the L2 loss of the projected 2D joints J_b, and L_{prior} represents the Mixture Gaussian prior loss of the SMPL parameters, which supervises the plausibility of 3D joint rotation and body shape. The terms $w(\cdot)$ denote the corresponding loss weights.

Root Depth Loss. Similar to [40], the losses incorporated for supervising depth learning consist of cross-entropy loss on the estimated bins B and L1 loss on the bin index b. These losses are defined as follows:

$$\mathcal{L}_{bins} = -\sum_{i=0}^{N-1} B_i^{GT} \log B_i^{pred} \tag{10}$$

$$\mathcal{L}_{id} = |b^{GT} - b^{pred}| \tag{11}$$

Here we need to generate ground truth bins in order to calculate the $\mathcal{L}_{bins}$ loss. Firstly, we compute the bin index using Eq. 6. Then the procedure to generate the final bins from the index is outlined in the pseudo code of Algorithm 1, which is self-explainable.

Algorithm 1. Generate *Ground truth Bins*.

$N \leftarrow$ Number of *Bins*
$b(\hat{d}) \leftarrow$ Bin index
$Arange(x) \leftarrow$ List of integers from 0 to x
$ABS(x) \leftarrow$ Absolute value of x
$Clip(x, y, z) \leftarrow$ Clip each value of x between y and z
$\mathit{GTBins} = 1 - \text{Clip}\left(\text{ABS}\left(\text{Arange}(N) - b(\hat{d})\right), 0, 1\right)$

The expression for the depth loss is as follows:

$$\mathcal{L}_d = w_{bins}\mathcal{L}_{bins} + w_{id}\mathcal{L}_{id} \tag{12}$$

Again $w(\cdot)$ denotes the corresponding loss weights.

Focal Length Estimation. In order to estimate the focal length effectively, we design a hybrid network called TransFocal (Fig. 4), which combines a convolutional neural network (CNN) with a vision transformer. CNNs have a stronger inductive bias compared to transformer networks when it comes to processing images [4]. This allows them to learn embeddings quickly from a smaller subset of data. However, when trained on a sufficient amount of data, vision transformers can outperform similar state-of-the-art CNN models [10] due to their self-attention architectures that offer better generalization properties [4]. Recent studies have shown that combining CNN embeddings with vision transformers results in a hybrid system that performs better than larger and deeper

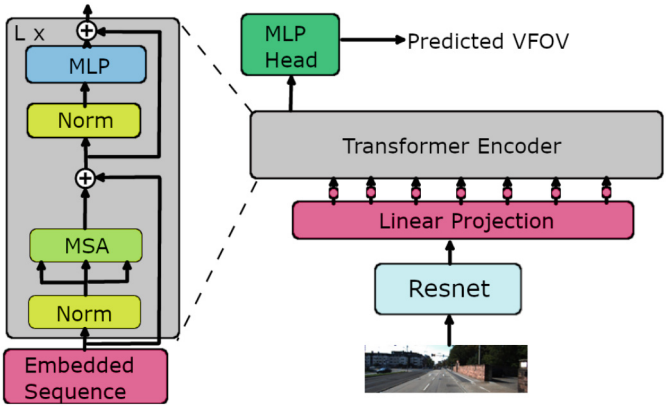

Fig. 4. TransFocal Architecture. The image is input into a ResNet backbone to create embeddings which are then projected into latent embedding space and used as input for the vision transformer. The output from the transformer's many layers is then decoded using a fully connected layer. This figure is adopted from the paper we presented at VISAPP 2023 [1].

vision transformers, with significantly lower computational cost for fine-tuning [11]. By leveraging the strengths of both architectures, we can train the model on less data while achieving improved accuracy.

Since the focal length measured in pixels has an unbounded range and changes when resizing images, we instead estimate the vertical field of view (vfov) v in radians and then convert it to the focal length f_y using the following equation:

$$f_y = \frac{0.5h}{\tan(0.5v)} \tag{13}$$

In the above equation, h represents the image height measured in pixels. We follow the assumptions made in [63] and [30], where we consider zero camera yaw and assume that the effective focal length values are the same in both directions, i.e., $f_x = f_y = f$.

TransFocal takes a complete image as input to predict its vfov, which remains the same for all subjects in the image of a video sequence. This means that inference needs to be performed only once to obtain absolute coordinates for each frame of the video sequence. The full image contains rich cues that facilitate transformer's self-attention. In particular, vanishing points and geometric lines help the network semantically reason about the vertical field of view of an image.

Using a bin technique similar to our absolute depth map head and the approach in [30], we discretize the vfov space v into B bins, effectively transforming the harder regression problem into an easier classification problem. Additionally, similar to our depth map head, we aggregate the predicted probability mass using a soft argmax operation. During our testing, we found that combining the cross-entropy loss $\mathcal{L}_{CE}$ (with a smaller weight) and the softargmax-biased L2 loss [30] $\mathcal{L}_{agmax}$ improves model convergence. Therefore, the final loss $\mathcal{L}_{foc}$ is defined as:

$$\mathcal{L}_{foc} = \lambda_{agmax}\mathcal{L}_{agmax} + \lambda_{CE}\mathcal{L}_{CE} \tag{14}$$

In the equation above, λ_{agmax} and λ_{CE} represent the weights associated with the respective loss terms. This formulation helps the model effectively estimate the focal length based on the vertical field of view prediction.

4 System Configurations and Implementation Details

In the following, we will describe the system configurations the implementation details of both Absolute-ROMP (especially the absolute depth map head), and TransFocal (i.e., the focal length estimation network). We will cover important topics including network architectures, training datasets, training setting details, and evaluation metrics.

4.1 Absolute-ROMP

Network Architecture. First HRNet-32 [56] is employed as the backbone for Absolute-ROMP. CoordConv [43] is also maintained from ROMP to enhance the spatial information. In this way, the feature set extracted by the backbone is the combination of a coordinate index map and output feature embeddings. This feature set is then used as input to the four heads of Absolute-ROMP for complete end to end prediction: the Body Center Map head, the Camera Map head, the SMPL Map head and our newly-designed Absolute Root Depth Map. The architecture of the Absolute Root Depth Map is similar to the other map heads. As in, it uses a series of feed forward ResNet blocks after a Trans block. Finally a 1×1 convolution layer is used to get the final output map. For details on the Trans block and basic ResNet block architecture of the map heads please refer to [55]. The only alteration made for the Absolute Root Depth Map is the output of the final 1×1 convolutional layer, with a size of $120 \times 64 \times 64$: the binning resolution is set to 120 and the map size is 64×64.

Training Datasets. The basic training datasets used in the experiments include three 3D pose datasets and four in-the-wild 2D pose datasets. The three 3D pose datasets are Human3.6M [18], MPI-INF-3DHP [48] and MuCo-3DHP [48]), and the four in-the-wild 2D pose datasets are MS COCO [42], MPII [3], LSP [20] and Crowdpose [35]. Pseudo 3D annotations from [31] and pseudo 3D labels of 2D pose datasets provided by [22] are also used. 3D datasets provide us with ground truth depth shape and 3D pose. Whereas 2D datasets are added to increase training data and improve generality where the projection of the 3d joints back to 2D is compared to ground truth. Psuedo 3D data acts as an in between. We also use the 3DPW [47] training set for fine tuning the Absolute-ROMP model only for evaluation on a 3D pose dataset 3DPW.

Training Setting Details. During training, the input images are resized to 512×512. For keeping the same aspect ratio images with a different aspect ratio are padded with zeros. The size of the backbone feature is $H_b = W_b = 128$. The maximum number of detection of persons is $N = 64$. The learning rate used is 5e-5. The batch size is set to 26. We adopt the Adam optimizer [26] for training, and train the model until performance plateaus on the validation set.

Evaluation Benchmarks and Metrics. Our trained model is evaluated on the Human3.6M [18], MuPoTS-3D [49] and 3DPW [47]. To evaluate the 3D pose accuracy, both mean per joint position error (MPJPE) [21] and Procrustes-aligned MPJPE (PMPJPE) are employed. MPJPE measures the average Euclidean distance between the location of real-life joints on human bodies and the locations of predicted joints on 3D poses after translating the root joints ('pelvises') of estimated bodies to the ground-truth root. Procrustes-aligned MPJPE, on the other hand, uses Procrustes' alignment (PA) [46] to solve for translation, scale and rotation between the estimated bodies and the ground truth and thus mostly focuses on the pose error.

For root depth estimation, Mean root position error ($MRPE_z$) [50] and 3D percentage of correct absolute keypoints (PCK_{abs}) [50] are employed. $MRPE$ is the mean of the euclidean distance between the estimated coordinates of the predicted absolute root and ground truth absolute root, and $MRPE_z$ is measure of correctness of the depth as it only looks at the z-axis. On the other hand, PCK_{abs}, the 3D percentage of correct absolute keypoints, treats a joint's absolute prediction as correct if it lies within a 15cm from the ground truth joint location.

4.2 TransFocal

Network Architecture. The architecture has been shown in Fig. 4. Similar to [60], we use ResNet [16] as the CNN backbone. Then learnable patch embeddings are applied to patches extracted from the ResNet output. Each patch embedding's kernel size is equal to the patch size, in that the input sequence is obtained by simply flattening the spatial dimensions of the ResNet features and projecting to the dimension of the Vision Transformer.

For completion, We also show the overview of the transformer encoder architecture in Fig. 4, to help with the explanation stated below. The input of the first Transformer layer z_0 is calculated as follow:

$$z_0 = l^1E; l^2E; l^3E....; l^nE \tag{15}$$

where z_0 is mapped into a latent n-dimensional embedding space using a trainable linear projection layer and E is the patch embedding projection. These patches are then fed into the vision transformer, specifically the VIT-B16 [10] variant. There are L Transformer layers which consist of multi-headed self-attention (MSA) and multi-layer perceptron (MLP) blocks. At each transformer layer ℓ, the input of the self-attention block is a triplet of Q (query), K (key), and V (value). They are computed from the output of the previous layer by matrix multiplication with learnable parameters of weight matrices. The self-attention in the attention head AH is calculated as:

$$AH = softmax(\frac{Q \times K^T}{\sqrt{d}}) \cdot V \tag{16}$$

where d is the dimension of self-attention block. Multi-headed self-attention (MSA) means the attention head will be calculated m times by independent weight matrices, as $AH_1, AH_2, \cdots, AH_m$. The final $MSA(z_{\ell-1})$ is defined as:

$$MSA(z_{\ell-1}) = z_{\ell-1} + concat(AH_1; AH_2; ...; AH_m) \times W_o, \tag{17}$$

The output of MSA is then transformed by an MLP block with residual skip connection as the layer output as:

$$z_l = MLP(Norm(MSA(z_{\ell-1}))) + MSA(z_{\ell-1}) \tag{18}$$

where $Norm$ means the layer normalization operator. Finally the MLP Head, a fully connected layer, is used to decode the output of the vision transformer to obtain the predicted VFOV (Fig. 4). Going off the VIT transformer architecture [10], MLP layer is sufficient as a decoder in vision related tasks.

Training and Evaluation Datasets. For training the TransFocal model, a dataset is created using the Pano360 dataset [30]. Pano360 consists of real panoramic images taken from flickr [14] as well as synthetic panoramas. Due to a portion of flickr images being inaccessible, we were unable to recreate the complete dataset. The dataset is split for training and evaluation purposes. The code that is used for creating the dataset is available at [29]. Note that since the image generation parameters are randomized, it is difficult to recreate exact the same dataset.

Training Setting Details. The TransFocal model is trained with images of varied resolutions. The learning rate used is 1e-4. The weight decay is set to 1e-2. The batch size is set to 4. We adopt the Adam optimizer [26] for training. We train the model until performance plateaus on validation set.

5 Experimental Results and Discussions

In this section we report some performance comparison results on multiple datasets, and a ablation study in using the Absolute-ROMP. We also provide an analysis of the real-time performance of all the components, including the four head networks.

5.1 Real-Time Performance Analysis

In the following, we analyze the numbers of parameters of the four heads, and their inference times. The information shown in Table 1 is obtained on a Nvidia Tesla V100 GPU. Overall, the numbers of the parameters are around 170K, and the interference times for HRNET-32 feature output are from 0.7 ms to 0.8 ms, which ensures a real-time performance.

The inference times for all the components are also analyzed, as shown in Table 2. In particular, TransFocal, which is run on a GTX 1080ti at 25 ms per image, can be run in parallel with the Absolute ROMP network, making its inference time not affect the total run time, since the backbone HRNet-32 dominates the computation time (at 35 ms per image), which is 50 times more than each of the four heads. Adding all together, the inference time is still near real-time, at $\sim$ 23 fps on the Nvidia Tesla V100 GPU.

Table 1. Parameters and inference time on Tesla V100 for four heads.

Head	Parameters	Inference Time (ms)
Body Center	167,809	0.716
Camera	167,939	0.702
SMPL	176,974	0.767
Root Depth	175,544	0.700

Table 2. Inference time on Tesla V100 for every component except TransFocal which is run in parallel on a separate GPU(gtx1080ti).

Head	Inference Time (ms)
Backbone(HRNET-32)	35
Body Center	0.716
Camera	0.702
SMPL	0.767
Root Depth	0.700
TransFocal	28(GTX 1080ti)
SMPL Wrapper	0.1
output matching	0.01

5.2 Performance Evaluation and Comparison

In the following, we evaluate the performance of Absolute-ROMP against the state-of-the-art (SOTA) methods on the Human3.6M, MuPoTS-3D and 3DPW datasets. We evaluate two indicators - $MRPE_z$ and PCK_{abs} (for absolute depth estimation) on the first two datasets and two performance indicators for joint positioning after root alignment - $MJPJPE$ and $PMPJPE$ on the 3DPW dataset. Since ROMP [55] does not have absolute depth prediction, it is only evaluated on the 3DPW dataset. We also show importance of absolute positioning by qualitatively comparing Absolute-ROMP and ROMP on the MuPoTS-3D dataset.

Performance Evaluation on Absolute Depth Prediction. Table 3 shows the root joint localization results on Human3.6M dataset. The three baselines reported in the first 3 rows all follow a two-stage approach [50], where 2D pose and 3D pose are estimated separately, and then an optimization process is adopted to obtain the global root joint location that minimizes the re-projection error. The baseline "w/o limb joints" refers to optimization using only head and body trunk joints. The baseline "with RANSAC" refers to randomly sampling the set of joints used for optimization with RANSAC. The baseline results are taken from [50].

The Absolute-ROMP is compared with two state-of-the-art (SOTA) approaches [40, 50] as well. In [50] a multi-stage approach is used, whereas in [40] a graph convolution network model is used. For a fair comparison, we used the baseline focal length (e.t. the

Table 3. Comparison of $MRPE_z$ results with state-of-the-art (the lower, the better) on the Human3.6M dataset (The results with the baseline focal length is from our conference paper [1]).

Methods	Baseline	Base w/o limb joints	Base w RANSAC	RootNet [50]	HDNET [40]	Ours
$MRPE_z \downarrow$ with baseline focal length	261.9	220.2	207.1	108.1	69.9	**68.0**
$MRPE_z \downarrow$ with CamCalib [30]	-	-	-	-	-	**84.9**

Table 4. Comparison of PCK_{abs} results with SOTA methods (the higher, the better) on the MuPoTS-3D dataset (The results with the baseline focal length is from our conference paper [1]).

Methods	RootNet [50]	HDNET [40]	SMAP [62]	Ours
$PCK_{abs} \uparrow$ with baseline focal length	31.9	35.2	**35.4**	35.3
$PCK_{abs} \uparrow$ with CamCalib [30]	-	-	-	30.9

Table 5. Comparisons to the state-of-the-art methods (the lower, the better) on 3DPW (from our conference paper [1]).

Methods	YOLO + VIBE [27]	ROMP(HRNET-32) [55]	Absolute-ROMP(HRNET-32)
MPJPE$\downarrow$	82.9	**76.7**	84.0
PMPJPE $\downarrow$	51.9	**47.3**	50.5

ground truth) when generating our Absolute-ROMP results, in the same way as in the SOTA approaches. Our model is able to outperform these two SOTA approaches while maintaining a real-time inference. As we described before, our system runs at $\sim$ 23 fps on a Nvidia Tesla V100 GPU. Our root joint localization head achieves a 68.0 mm accuracy with the baseline focal length information. Our end to end architecture is able to look at the big picture, picking up distance cues in the background. While this gives us an upper bound for the Absolute-ROMP accuracy (since the focal length is accurate), we also test the Absolute-ROMP with a state-of-the-art (SOTA) approach CamCalib [30] for estimating the focal length, which gives us a lower bound of the accuracy using Absolute-ROMP. Due to the limitation of our computational facilities, we have not put Absolute-ROMP and TransFocal together into one system. Nevertheless, in the next experiment on TransFocal testing, we can see that Transfocal outperforms CamCalib [30] so we can expect a performance improvement when using the TransFocal for focal length estimation.

We showcase the 3D PCK_{abs} performance of Absolute-ROMP on the MuPoTS-3D dataset in Table 4. Our model has comparable performance to the SOTA methods including RootNet [50], HDNET [40] and SMAP [62], in terms of the 3D percentage of correct absolute keypoints, all using the baseline focal length information. The reason for slightly better performance of SMAP [62] than ours is because an additional network called RefineNet is employed in SMAP to further refine the output from the initial network, thus filling in missing body parts and improving the visible ones. However, this technique works on a 3D prior and might not function well if test scenario is very different from the training data. Again, using the focal length estimated with a SOTA approach CamCalib [30] hurts the performance due to missing information of

the ground truth focal length but this allows us to use the model when the focal length is unavailable as is the case in most images available online.

Performance Evaluation Against Methods Without Absolute Depth Prediction. We also compare MJPJPE and PMPJPE performance on the 3DPW dataset in Table 5 with SOTA methods (YOLO + VIBE [27] and ROMP(HRNET-32) [55]), which do not predict absolute depth, as MJPJPE and PMPJPE are evaluated right after root alignment. Even with added depth map prediction in our Absolute-ROMP, we are able to maintain comparable performance with the SOTA. Note that in order to obtain abosulte depth information, the backbone in our approach has to compensate for the additional prediction of the root depth with a slightly higher joint error. This issue might be resolved by using a larger backbone (such as HRNET-48) even though that would inevitably slow down inference times. Nevertheless, a qualitative comparison of Absolute-ROMP with ROMP on the MuPoTS-3D dataset in Fig. 1 highlights the importance of absolute global coordinates. Thanks to absolute depth information while positioning the meshes, we improve the location accuracy therefore correctly placing people hugging each other.

Performance Evaluation of TransFocal. Finally, we present the results of our proposed focal length estimation method, TransFocal, and compare its performance against a state-of-the-art (SOTA) approach CamCalib [30]. We have conducted an evaluation, the findings of which are summarized in Table 6.

Note that our TransFocal model has been trained on a carefully curated dataset composed of partially available images from the Pano360 dataset. This dataset was constructed to include a diverse range of camera viewpoints from panoramic scenes. In contrast, the CamCalib model provided by the author [30] was pretrained on a dataset created from the complete Pano360 dataset. To ensure a fair and unbiased comparison, we evaluate both TransFocal and CamCalib on a subset of the dataset that was unseen by our model during the training phase. This approach allows us to assess the generalization and robustness of our method in handling new, unseen panoramic images. The results obtained from our experiments demonstrate the superiority of the TransFocal model over CamCalib. In fact, TransFocal consistently outperforms CamCalib by up to an impressive 40%. These findings highlight the effectiveness of our proposed approach and its ability to accurately calibrate the camera parameters. These results validate the efficacy of our method in solving the challenging task of camera calibration, even with limited data availability.

Table 6. vfov error results comparison with state-of-the-art on dataset created from Pano360 dataset (from our conference paper [1]).

Methods	CamCalib [30]	TransFocal
vfov diff(degrees)	26.35	**15.59**

Table 7. vfov error comparison after 1 epoch with different losses for supervision (from our conference paper [1]).

Methods	softargmax-biased-L2	softargmax-biased-L2+cross entropy
vfov diff	15.8	**15.6**

Table 8. MRPE comparison after 1k iterations with different bin resolutions.

Methods	$MRPE_z$(mm)
60	248
120	239
240	225

Table 9. MRPE comparison after 1k iterations binning vs numerical output.

Methods	$MRPE_z$(mm)
binning(120)	239
numerical	276

5.3 Ablation Study

Here we show the improvement in performance when we use a combination loss instead of just employing the Softargmax-biased-L2 loss when training TransFocal. We report mean error after training for 1 epoch while using Softargmax-biased-L2 loss alone and with cross entropy loss in Table 7. This indicates that the cross entropy loss acts as a guide for the gradient descent direction when the model is starting out, thus adding to the speed of convergence of the model.

We also examine how different binning resolution effects output, as shown in Table 8. When it comes to bin size it seems the highest resolution should be chosen that would result in reasonable memory usage and inference times. However a larger bin size, i.e. higher bin resolution, is preferred to account for a larger interval. Typically, in our testing a factor of 10 i.e. ten times the interval of the focal length that is to be estimated, is optimal for balancing inference time and memory usage when compared with the distance in meters. The reason for this is that the gap between different bin sizes (as we go higher than factor of 10) becomes almost negligible as we keep training further. In our case, we chose a bin resolution of 120 as our prediction interval was 12 m.

Furthermore, we compare how numerical output compares to softmax binning, as shown in Table 9. The results confirm superiority of soft max binning when compared to unbounded numeric input.

6 Conclusion and Future Work

In conclusion, our Absolute-ROMP is built upon an end to end one-stage network ROMP for monocular multi-person 3D mesh regression from a single RGB image, by

adding in absolute distance prediction. To achieve this, we introduce absolute root joint localization in the camera coordinate frame, which enables the estimation of 3D mesh coordinates of all persons in the image and their root joint locations normalized by the focal length. Further, for eliminating the need for known intrinsic parameters of the camera, we design and train a focal length prediction network called TransFocal, which is a CNN + Transformer hybrid model. We evaluate Absolute-ROMP on the root joint localization and root-relative 3D pose estimation tasks on publicly available multi-person 3D pose datasets. We also evaluate TransFocal on a dataset created from the Pano360 dataset. Quantitative and qualitative results show that our approach outperforms or has competitive performance to the state-of-the-art approaches on multiple benchmark datasets, under various evaluation metrics. Additionally, our method is applicable to in-the-wild images and videos due to its real-time performance.

Future work include the following several directions. First, for the core algorithms, incorporating the absolute camera parameters would eliminate the need for predicting the depth separately. This would require the use of accurate absolute multi-person 3D datasets in a variety of scenarios, such as the synthetic dataset AGORA [51]. Second, real-time clothes and texture prediction on top of the multi-person 3D mesh regression would be especially beneficial for both virtual reality and augmented reality applications with realistic 3D rendering. Finally, incorporating labels for children would improve the absolute location prediction for all, as children possess different body proportions when compared with adults on average.

References

1. Abdulrahman, B., Zhu, Z.: Absolute-romp: absolute multi-person 3D mesh prediction from a single image. In: Radeva, P., Farinella, G.M., Bouatouch, K. (eds.) Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2023, Volume 5: VISAPP, Lisbon, Portugal, 19–21 February 2023, pp. 69–79. SCITEPRESS (2023). https://doi.org/10.5220/0011629500003417
2. Aitpayev, K., Gaber, J.: Creation of 3D human avatar using kinect. Asian Trans. Fundam. Electron. Commun. Multimedia **1**(5), 12–24 (2020)
3. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: new benchmark and state of the art analysis. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3686–3693 (2014). https://doi.org/10.1109/CVPR.2014.471
4. Bai, Y., Mei, J., Yuille, A., Xie, C.: Are transformers more robust than CNNs? In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (2021). https://openreview.net/forum?id=hbHkvGBZB9
5. Benzine, A., Chabot, F., Luvison, B., Pham, Q.C., Achrd, C.: Pandanet: anchor-based single-shot multi-person 3D pose estimation. arXiv (2021). https://doi.org/10.48550/ARXIV.2101.02471. arXiv:2101.02471
6. Boulic, R., Bécheiraz, P., Emering, L., Thalmann, D.: Integration of motion control techniques for virtual human and avatar real-time animation, pp. 111–118 (1997). https://doi.org/10.1145/261135.261156
7. Carbonera Luvizon, D., Tabia, H., Picard, D.: Learning features combination for human action recognition from skeleton sequences. Pattern Recognit. Lett. **99**, 13–20 (2017). https://doi.org/10.1016/j.patrec.2017.02.001. https://www.sciencedirect.com/science/article/pii/S0167865517300326. User Profiling and Behavior Adaptation for Human-Robot Interaction

8. Choi, H., Moon, G., Lee, K.M.: Pose2mesh: graph convolutional network for 3D human pose and mesh recovery from a 2D human pose (2020). https://doi.org/10.48550/ARXIV.2008.09047. arXiv:2008.09047
9. Dabral, R., Gundavarapu, N.B., Mitra, R., Sharma, A., Ramakrishnan, G., Jain, A.: Multi-person 3D human pose estimation from monocular images (2019). https://doi.org/10.48550/ARXIV.1909.10854. arXiv:1909.10854
10. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale (2020). https://doi.org/10.48550/ARXIV.2010.11929. arXiv:2010.11929
11. Dosovitskiy, A., et al.: vision transformer github repository. GitHub repository (2020)
12. Du, G., Zhang, P.: Markerless human-robot interface for dual robot manipulators using kinect sensor. Robot. Comput. Integr. Manuf. **30**(2), 150–159 (2014). https://doi.org/10.1016/j.rcim.2013.09.003. https://www.sciencedirect.com/science/article/pii/S0736584513000628
13. Fang, H.S., Cao, J., Tai, Y.W., Lu, C.: Pairwise body-part attention for recognizing human-object interactions (2018). https://doi.org/10.48550/ARXIV.1807.10889. arXiv:1807.10889
14. Flickr: Yahoo (2022). https://www.flickr.com/
15. Grompone von Gioi, R., Jakubowicz, J., Morel, J.M., Randall, G.: LSD: a fast line segment detector with a false detection control. IEEE Trans. Pattern Anal. Mach. Intell. **32**(4), 722–732 (2010). https://doi.org/10.1109/TPAMI.2008.300
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015). https://doi.org/10.48550/ARXIV.1512.03385. arXiv:1512.03385
17. Hold-Geoffroy, Y., et al.: A perceptual measure for deep single image camera calibration (2017). https://doi.org/10.48550/ARXIV.1712.01259. arXiv:1712.01259
18. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2014). https://doi.org/10.1109/TPAMI.2013.248
19. Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image (2020). https://doi.org/10.48550/ARXIV.2006.08586. arXiv:2006.08586
20. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: Proceedings of the British Machine Vision Conference, pp. 12.1–12.11. BMVA Press (2010). https://doi.org/10.5244/C.24.12
21. Joo, H., et al.: Panoptic studio: a massively multiview system for social motion capture. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 3334–3342 (2015). https://doi.org/10.1109/ICCV.2015.381
22. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3D human model fitting towards in-the-wild 3D human pose estimation (2020). https://doi.org/10.48550/ARXIV.2004.03686. arXiv:2004.03686
23. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose (2017). https://doi.org/10.48550/ARXIV.1712.06584. arXiv:1712.06584
24. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3D human dynamics from video (2018). https://doi.org/10.48550/ARXIV.1812.01601. arXiv:1812.01601
25. Kendall, A., Grimes, M., Cipolla, R.: Posenet: a convolutional network for real-time 6-DOF camera relocalization. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2938–2946 (2015). https://doi.org/10.1109/ICCV.2015.336
26. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2014). https://doi.org/10.48550/ARXIV.1412.6980. arXiv:1412.6980
27. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation (2019). https://doi.org/10.48550/ARXIV.1912.05656. arXiv:1912.05656
28. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: Pare: part attention regressor for 3D human body estimation (2021). https://doi.org/10.48550/ARXIV.2104.08527. arXiv:2104.08527

29. Kocabas, M., Huang, C.H.P., Tesch, J., Müller, L., Hilliges, O., Black, M.J.: Spec github repository. GitHub repository (2021)
30. Kocabas, M., Huang, C.H.P., Tesch, J., Müller, L., Hilliges, O., Black, M.J.: Spec: seeing people in the wild with an estimated camera (2021). https://doi.org/10.48550/ARXIV.2110.00620. arXiv:2110.00620
31. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop (2019). https://doi.org/10.48550/ARXIV.1909.12828. arXiv:1909.12828
32. Lee, J.H., Kim, C.S.: Monocular depth estimation using relative depth maps, pp. 9721–9730 (2019). https://doi.org/10.1109/CVPR.2019.00996
33. Lee, J., Go, H., Lee, H., Cho, S., Sung, M., Kim, J.: CTRL-C: camera calibration transformer with line-classification (2021). https://doi.org/10.48550/ARXIV.2109.02259. arXiv:2109.02259
34. Li, J., Wang, C., Liu, W., Qian, C., Lu, C.: Hmor: hierarchical multi-person ordinal relations for monocular multi-person 3D pose estimation (2020). https://doi.org/10.48550/ARXIV.2008.00206. arXiv:2008.00206
35. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: efficient crowded scenes pose estimation and a new benchmark (2018). https://doi.org/10.48550/ARXIV.1812.00324. arXiv:1812.00324
36. Li, Y.L., et al.: Detailed 2D-3D joint representation for human-object interaction (2020). https://doi.org/10.48550/ARXIV.2004.08154. arXiv:2004.08154
37. Li, Y.L., et al.: Pastanet: toward human activity knowledge engine (2020). https://doi.org/10.48550/ARXIV.2004.00945. arXiv:2004.00945
38. Li, Z., et al.: Learning the depths of moving people by watching frozen people (2019). https://doi.org/10.48550/ARXIV.1904.11111. arXiv:1904.11111
39. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos (2018). https://doi.org/10.48550/ARXIV.1804.00607. arXiv:1804.00607
40. Lin, J., Lee, G.H.: Hdnet: human depth estimation for multi-person camera-space localization (2020). https://doi.org/10.48550/ARXIV.2007.08943. arXiv:2007.08943
41. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection (2018)
42. Lin, T.Y., et al.: Microsoft coco: common objects in context (2014). https://doi.org/10.48550/ARXIV.1405.0312. arXiv:1405.0312
43. Liu, R., et al.: An intriguing failing of convolutional neural networks and the coordconv solution (2018). https://doi.org/10.48550/ARXIV.1807.03247. arXiv:1807.03247
44. Lo Presti, L., La Cascia, M.: 3D skeleton-based human action classification: a survey. Pattern Recogn. **53**, 130–147 (2016). https://doi.org/10.1016/j.patcog.2015.11.019. https://www.sciencedirect.com/science/article/pii/S0031320315004392
45. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.: Smpl: a skinned multi-person linear model, vol. 34 (2015). https://doi.org/10.1145/2816795.2818013
46. Luo, B., Hancock, E.R.: Feature matching with procrustes alignment and graph editing. In: Image Processing and Its Applications, 1999. Seventh International Conference on (Conf. Publ. No. 465), vol. 1, pp. 72–76 (1999). https://doi.org/10.1049/cp:19990284
47. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In: European Conference on Computer Vision (ECCV) (2018)
48. Mehta, D., et al.: Monocular 3D human pose estimation in the wild using improved CNN supervision (2016). https://doi.org/10.48550/ARXIV.1611.09813. arXiv:1611.09813
49. Mehta, D., et al.: Single-shot multi-person 3D pose estimation from monocular RGB (2017). https://doi.org/10.48550/ARXIV.1712.03453. arXiv:1712.03453

50. Moon, G., Chang, J.Y., Lee, K.M.: Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image (2019). https://doi.org/10.48550/ARXIV.1907.11346. arXiv:1907.11346
51. Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J.: AGORA: avatars in geography optimized for regression analysis. In: Proceedings IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
52. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.C.: Learning human-object interactions by graph parsing neural networks (2018). https://doi.org/10.48550/ARXIV.1808.07962. arXiv:1808.07962
53. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks (2015). https://doi.org/10.48550/ARXIV.1506.01497. arXiv:1506.01497
54. Rogez, G., Weinzaepfel, P., Schmid, C.: LCR-net+: multi-person 2D and 3D pose detection in natural images. IEEE Trans. Pattern Anal. Mach. Intell. **42**, 1146–1161 (2019). https://doi.org/10.1109/tpami.2019.2892985
55. Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, one-stage, regression of multiple 3d people (2020). https://doi.org/10.48550/ARXIV.2008.12272. arXiv:2008.12272
56. Wang, J., et al.: Deep high-resolution representation learning for visual recognition (2019). https://doi.org/10.48550/ARXIV.1908.07919. arXiv:1908.07919
57. Workman, S., Greenwell, C., Zhai, M., Baltenberger, R., Jacobs, N.: Deepfocal: a method for direct focal length estimation. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 1369–1373 (2015). https://doi.org/10.1109/ICIP.2015.7351024
58. Workman, S., Zhai, M., Jacobs, N.: Horizon lines in the wild (2016). https://doi.org/10.48550/ARXIV.1604.02129. arXiv:1604.02129
59. Xu, Y., Zhu, S.C., Tung, T.: Denserac: joint 3D pose and shape estimation by dense render-and-compare (2019). https://doi.org/10.48550/ARXIV.1910.00116. arXiv:1910.00116
60. Yang, G., Tang, H., Ding, M., Sebe, N., Ricci, E.: Transformer-based attention networks for continuous pixel-wise prediction (2021). https://doi.org/10.48550/ARXIV.2103.12091. arXiv:2103.12091
61. Zanfir, A., Marinoiu, E., Zanfir, M., Popa, A.I., Sminchisescu, C.: Deep network for the integrated 3D sensing of multiple people in natural images. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc. (2018). https://proceedings.neurips.cc/paper/2018/file/6a6610feab86a1f294dbbf5855c74af9-Paper.pdf
62. Zhen, J., et al.: Smap: single-shot multi-person absolute 3D pose estimation (2020). https://doi.org/10.48550/ARXIV.2008.11469. arXiv:2008.11469
63. Zhu, R., et al.: Single view metrology in the wild (2020). https://doi.org/10.48550/ARXIV.2007.09529. arXiv:2007.09529
64. Zimmermann, C., Welschehold, T., Dornhege, C., Burgard, W., Brox, T.: 3D human pose estimation in RGBD images for robotic task learning (2018). https://doi.org/10.48550/ARXIV.1803.02622. arXiv:1803.02622