

# Absolute-ROMP: Absolute Multi-Person 3D Mesh Prediction from a Single Image

Bilal Abdulrahman<sup>1</sup> <sup>a</sup> and Zhigang Zhu<sup>2</sup> <sup>b</sup>

<sup>1</sup>The CUNY Graduate Center, New York, NY 10016, U.S.A.

<sup>2</sup>The CUNY City College and Graduate Center, New York, NY 10031, U.S.A.

**Keywords:** Machine Learning, Computer Vision, 3D reconstruction, Camera Calibration, Mesh Regression, Pose Prediction, Human Mesh Regression.

**Abstract:** Recovering multi-person 3D poses and shapes with absolute scales from a single RGB image is a challenging task due to the inherent depth and scale ambiguity from a single view. Current works on 3D 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 defined as the pelvis joint. In this paper, we build upon an existing multi-person 3D mesh predictor network, ROMP, to create Absolute-ROMP. By adding absolute root joint localization in the camera coordinate frame, we are able to estimate multi-person 3D poses and shapes with absolute scales from a single RGB image. Such a single-shot approach allows the system to better learn and reason about the inter-person depth relationship, thus improving multi-person 3D estimation. In addition to this end to end network, we also train a CNN and transformer hybrid network, called TransFocal, to predict the focal length of the image's camera. Absolute-ROMP estimates the 3D mesh coordinates of all persons in the image and their root joint locations normalized by the focal point. We then use TransFocal to obtain focal length and get absolute depth information of all joints in the camera coordinate frame. 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 evaluate TransFocal on dataset created from the Pano360 dataset and both are applicable to in-the-wild images and videos, due to real time performance.

## 1 INTRODUCTION

3D Human pose and shape estimation is one of the most active fields of research within the current landscape of computer vision and AI thanks to its many applications in robotics (Du and Zhang, 2014; Zimmermann et al., 2018), activity recognition (Lo Presti and La Cascia, 2016; Carbonera Luvizon et al., 2017), graphics (Boulic et al., 1997; Aitpayev and Gaber, 2020) and human-object interaction detection (Fang et al., 2018; Qi et al., 2018; Li et al., 2020b; Li et al., 2020c). Current works on 3D 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, defined as the pelvis joint. Due to the inherent depth and scale ambiguity of a single view, it is quite a challenge to accurately recover multi-person 3D pose and shape with absolute scales from a single

RGB image. Addressing this ambiguity requires assessing cues in the image as a whole, such as scene layouts, body dimensions, and inter-person relationships. This paper aims to address the problem of estimating absolute 3D pose and shape of multiple people simultaneously from a single RGB image. Compared 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).

Estimating the absolute 3D location of each person in an image is essential for understanding human-to-human interactions (Figure 3). Since multi-person activities take place in cluttered scenes, inherent depth ambiguity and occlusions make it challenging to estimate the absolute positions of multiple individuals. We argue that body dimensions alone paint a vague picture of absolute depth. Robust estimation of global positions requires information cues over the entire image, such as geometric cues, human body sizes in

<sup>a</sup>  <https://orcid.org/0000-0002-1164-1976>

<sup>b</sup>  <https://orcid.org/0000-0002-9990-1137>

the image, any occlusions which might affect the perceived sizes, layout of the entire scene etc.

Most existing methods for absolute multi-person 3D pose estimation extend the 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 the person from the cropped image (Moon et al., 2019) or compute it based on the prior knowledge about the body size (Dabral et al., 2019), which ignores the global context of the whole image. While others employ complicated architecture with extensive steps, drastically slowing down inference (Lin and Lee, 2020).

In this paper, we build upon ROMP (Sun et al., 2020), 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 call the revised network *Absolute-ROMP*. We adopt the same end to end pipeline 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 train a CNN and a vision transformer hybrid network to predict the vertical field of view of the image, which is used to estimate the focal length. With the predicted focal length we can estimate absolute distance in camera coordinates without the need for camera intrinsic parameters. Our model uses embeddings from ResNet (He et al., 2015), which are then converted to tokens to be fed into the vision transformer (Dosovitskiy et al., 2020a). The added attention from the transformer gives our network an edge over previous work (Kocabas et al., 2021c).

Our contributions in this work are:

- We propose an absolute depth estimation head for the ROMP network using a combination loss.
- We design and train TransFocal, a network to predict focal length of the image thus negating the requirement of intrinsic parameters.

- Quantitative and qualitative results show that our approach outperforms or has competitive performance to the state-of-the-art on multiple benchmark datasets under various evaluation metrics.

This paper is organized as 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

Parametric human body models allow complex 3D human mesh vertices to be encoded into low dimensional parameter vectors. These have been widely adopted since they allow regression of 3D meshes from images, such as the Skinned Multi-Person Linear Model (SMPL) (Loper et al., 2015). Various weak supervision techniques have led to reasonable accuracy for single-person 3D mesh regression with various techniques, such as those using semantic segmentation (Xu et al., 2019), geometric prior (Kanazawa et al., 2017), motion analysis (Kanazawa et al., 2018; Kocabas et al., 2019) and 2D pose (Choi et al., 2020). The work in (Kocabas et al., 2021a) uses a part-guided attention mechanism to overcome occlusions by exploiting information about the visibility of individual body parts while leveraging information from neighboring body-parts to predict occluded parts. Whereas (Kocabas et al., 2021c) uses predicted camera calibration parameters to aid in the regression of the body mesh parameters.

### 2.2 Multi-Person 3D Pose Estimation

For multi-person 3D pose estimation, (Mehta et al., 2017) proposes occlusion-robust pose-maps and exploits the body part association to avoid bounding box prediction. In (Benzine et al., 2021), an anchor-based one-stage model is proposed, which relies on a huge number of pre-defined anchor predictions and the positive anchor selection. To handle person-person occlusion, (Zhen et al., 2020) proposes a system that 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. Both (Rogez et al.,

2019) and (Moon et al., 2019) employ a top-down design, which estimates the target via regression from anchor-based feature proposals.

### 2.3 Multi-Person 3D Mesh Estimation

(Jiang et al., 2020) proposes a network for Coherent Reconstruction of Multiple Humans (CRMH). Built on Faster-RCNN (Ren et al., 2015), they use the RoI-aligned feature of each person to predict the SMPL parameters. (Zanfir et al., 2018b) estimates the 3D mesh of each person from its intermediate 3D pose estimation. (Zanfir et al., 2018a) further employs multiple scene constraints to optimize the multi-person 3D mesh results. All these methods follow a multi-stage design. The complex multi-step process requires a repeated feature extraction, which is computationally expensive. The ROMP (Sun et al., 2020), which our proposed model is based on, learns an explicit pixel-level representation with a holistic view, which improves accuracy in multi-person in-the-wild scenes.

### 2.4 Monocular Absolute Depth Estimation

Depth estimation from a single view suffers from inherent ambiguity. Nevertheless, several methods make remarkable advances in recent years (Li and Snavely, 2018; Lee and Kim, 2019). (Li et al., 2019) employs the use of mannequin challenge to obtain a dataset for depth estimation by employing the frozen poses and moving camera of the mannequin challenge. They generate training data using multi-view stereo reconstruction and adopt a data-driven approach to recover a dense depth map. However, such a depth map lacks scale consistency and cannot reflect the real depth. (Zhen et al., 2020) proposes a system that 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. (Lin and Lee, 2020) estimates the 2D human pose with heatmaps of the joints. Then these heatmaps are used as attention masks for pooling features from image regions corresponding to the target person. A skeleton-based Graph Neural Network (GNN) is used to predict the depth of each joint. (Li et al., 2020a) uses an integrated model to estimate human bounding boxes, human depths, and root-relative 3D poses simultaneously, with a coarse-to-fine architecture. All these methods either employ multi-step prediction or use large networks which slow down the inference. Our method is able to estimate depth and regress 3D mesh of multiple people in real time with high ac-

curacy while also regressing the shape parameters of individuals.

### 2.5 Focal Length Estimation

Recent works such as (Hold-Geoffroy et al., 2017; Kendall et al., 2015; Workman et al., 2016; Workman et al., 2015; Zhu et al., 2020) estimate camera parameters from a single image. To estimate camera rotations and fields of view, these methods train a neural network to leverage geometric cues in the image. (Workman et al., 2015) uses an AlexNet backbone to regress the horizontal field of view. Except (Workman et al., 2015), these methods discretize the continuous space of rotation into bins, casting the problem as a classification task and applying cross entropy (Workman et al., 2016) or KL-divergence (Hold-Geoffroy et al., 2017; Zhu et al., 2020) losses. (Kocabas et al., 2021c) also uses a binning technique. It trains a neural network with a bespoke-biased loss on a new collected dataset (Kocabas et al., 2021c). However none of these methods take advantage of the latest architecture innovation in the vision space, i.e., transformer networks (Dosovitskiy et al., 2020a). (Lee et al., 2021) takes both an image and line segments as input and regresses the camera parameters based on the transformer encode-decoder architecture. The line segments, extracted from the input image using the LSD algorithm (Grompone von Gioi et al., 2010), are also mapped to geometric tokens. However, 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 a simple MLP layer is used for decoding, thus an integration of a vision transformer and a CNN hybrid network with a combination of losses during supervision.

## 3 METHODOLOGY

Absolute-ROMP is an extension of the ROMP network (Sun et al., 2020), which regresses meshes in a one-stage fashion for multiple 3D people (thus termed ROMP). We will briefly explain the working of our Absolute-ROMP and, therefore by extension, of ROMP before going into details about our addition. Figure 1 shows overall system diagram. Absolute-ROMP employs a multi-head design with a HRNeT-32 backbone(Wang et al., 2019) and 4 head networks: Body Center Map, Camera Map, SMPL Map and Root Depth Map. Details of the backbone will be described in Section 4. Given a RGB image as in-

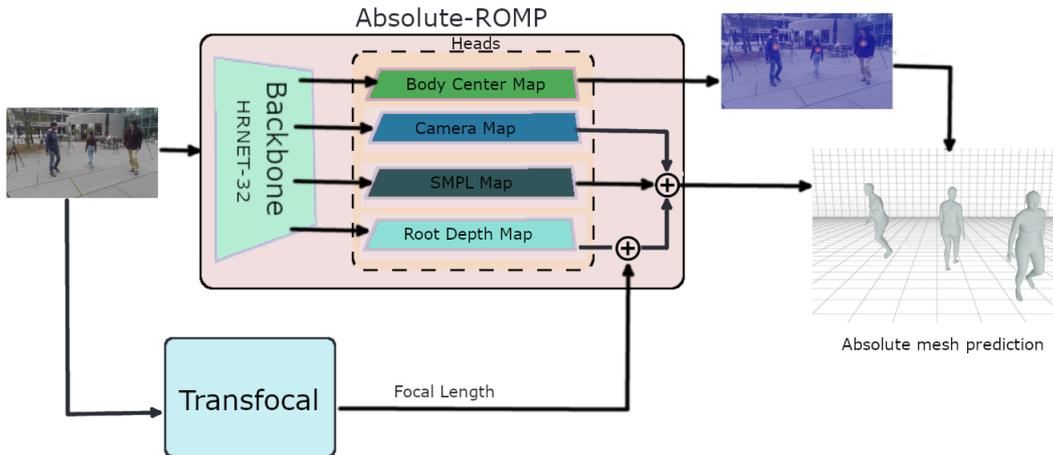


Figure 1: Overview of how the system works. Absolute-ROMP predicts the mesh parameters and depth. The focal length is predicted by TransFocal which are then used to get the complete absolute 3D coordinates.

put, it outputs a body center heatmap, camera parameters, SMPL parameters and an absolute depth map in the camera coordinate frame. The resolution for each map is  $64 \times 64$ . In the body center heatmap, Absolute-ROMP predicts the probability of each position being a human body center. Each body center is represented as a Gaussian distribution in the body center heatmap. At each position of the camera map, SMPL map and absolute depth map, it predicts the absolute depth of the root joint  $s_z$  (defined as the pelvis joint), camera parameters  $(t_x, t_y, s)$  (a weak-perspective camera model with translations in the x and y directions and a scale), and SMPL parameters (22×6 pose parameters and 10 shape parameters) of the person that takes the position as the center respectively.

During inference, we sample the 3D body mesh parameter results from the SMPL parameter map at the 2D body center locations parsed from the body center heatmap. We put the sampled parameters into the SMPL model to generate the 3D body meshes. We employ a weak-perspective camera model to project 3D body joints back to the 2D joints on the image plane. The reason for not employing perspective transform from the absolute coordinates for back projection is because the root depth map only predicts the absolute depth parameter  $s_z$ , not translation in the x or y direction. This is done in order to preserve real time inference, but it is not sufficient for a perspective projection back to the image. Weak-perspective camera model allows orthogonal back-projection of the 3D pose to 2D on the image, facilitating training the model with in-the-wild 2D pose datasets. This helps with robustness and generalization and also allows for more accurate position estimates than the body center heatmap, as the body center heatmap has limited resolution compared to the image. We will

provide more details on heatmap resolution in the next section. We employ the estimated focal length from TransFocal, the projected 2D pose and the predicted absolute depth of the root joint to get the absolute x and y coordinates of the root joint. Finally, using the 3D pose prediction from the SMPL mesh parameters, we infer the absolute location of each of the remaining joints.

### 3.1 Absolute Depth Map Head

From the original ROMP network (Sun et al., 2020), we maintain an output map size  $n \times H \times W$ , where  $n$  is the number of channels and  $H = W = 64$ , assuming that each location of the absolute depth map is the center of a human body, we estimate absolute depth of its corresponding root joint. Instead of directly regressing the numerical value of depth, we employ a binning technique in the log depth space. The binning resolution is set to 120. The range is chosen after taking prediction error mitigation and reasonable distance estimation based on all available data into consideration. Since different focal lengths of the camera affect the scales of a target person in the image, it is unrealistic to estimate the absolute depth from images taken by any arbitrary camera. 3D human datasets with absolute depth information have minimal variation in focal lengths (most datasets use the same camera for all images in each said dataset). This makes it challenging for the model to learn this variable and therefore can overfit on the focal lengths of the training datasets. The same person will have different dimensions in the images with cameras with different focal lengths. Therefore, the predicted depth is normalized (divided) by the focal length. The focal length is estimated by training another network Trans-

Focal, which we will detail in the next subsection.

We employ a soft-argmax operation on the bins for improved accuracy as compared to exact bin locations. Exact bins can only give integer outputs and therefore have limited precision. Whereas soft bins can output any number between bin indices depending on what number between 0 and 1 is the output of that bin. Therefore, there is minimal precision loss if any. This allows the actual prediction to be a ratio of the bins improving granularity and accuracy of the model. The bin index for the log depth space is computed as follows:

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

where  $b(\hat{d})$  is the bin index of the normalized depth  $\hat{d}$ .  $N$  is the total number of bins and  $[S, E]$  is the range of the bins. We can state this style of binning depth map as a collection of 1D heatmaps. There are  $64 \times 64$  predictions for each image. Therefore, we have a  $64 \times 64$  collection of 1D heatmaps. The predicted bin values  $B$  are converted back to normalized depth as follows:

$$\hat{d} = \exp\left[\frac{\sum_{i=0}^{N-1} B_i \times (\log E - \log S) + \log S}{N - 1}\right] \quad (2)$$

Similar to (Lin and Lee, 2020), the losses added to supervise depth learning are cross-entropy loss on the estimated bins  $B$  and L1 loss on the bin index  $b$  as follows:

$$\mathcal{L}_{bins} = - \sum_{i=0}^{N-1} B_i^{GT} \log B_i^{pred} \quad (3)$$

$$\mathcal{L}_{id} = |b^{GT} - b^{pred}| \quad (4)$$

To calculate  $\mathcal{L}_{bins}$  we have to generate ground truth bins. First, we compute bin index using equation 1. The procedure to generate the final bins from the index is then laid out in the pseudo code of Algorithm 1.

---

Algorithm 1: Generate *Ground truth Bins*.

---

- 1:  $N \leftarrow$  Number of *Bins*
  - 2:  $b(\hat{d}) \leftarrow$  Bin index
  - 3:  $Arrange(x) \leftarrow$  list of integers from 0 to  $x$
  - 4:  $ABS(x) \leftarrow$  Absolute value of  $x$
  - 5:  $Clip(x, y, z) \leftarrow$  Clip each value of  $x$  between  $(y, z)$
  - 6:  $GTBins = 1 - Clip(ABS(Arrange(N) - b(\hat{d})), 0, 1)$
- 

The final loss is as follows:

$$\begin{aligned} \mathcal{L}_{abs} = & \lambda_{pose} \mathcal{L}_{pose} + \lambda_{shape} \mathcal{L}_{shape} + \lambda_{j3d} \mathcal{L}_{j3d} \\ & + \lambda_{paj3d} \mathcal{L}_{paj3d} + \lambda_{pj2d} \mathcal{L}_{pj2d} \\ & + \lambda_{prior} \mathcal{L}_{prior} + \lambda_{bins} \mathcal{L}_{bins} + \lambda_{id} \mathcal{L}_{id} \end{aligned} \quad (5)$$

where  $\lambda_i$  represents the weight associated with the loss,  $\mathcal{L}_{pose}$  is the L2 loss of the pose parameters in the  $3 \times 3$  rotation matrix format,  $\mathcal{L}_{shape}$  is the L2 loss of the shape parameters,  $\mathcal{L}_{j3d}$  is the L2 loss of the 3D joints,  $\mathcal{L}_{paj3d}$  is the L2 loss of the 3D joints after Procrustes alignment,  $\mathcal{L}_{pj2d}$  is the L2 loss of the projected 2D joints, and  $\mathcal{L}_{prior}$  is the Mixture Gaussian prior loss of the SMPL parameters adopted in (Loper et al., 2015). please refer to (Sun et al., 2020) for further detail on these losses.

### 3.2 Focal Length Estimation

To estimate the focal length, we train a CNN and transformer hybrid network TransFocal. When it comes to images, convolutional neural networks by design have a stronger inductive bias compared to transformer networks (Bai et al., 2021). This results in them being able to learn to embeddings relatively quickly from a smaller subset of data. However, when trained on enough data, the vision transformer is able to outperform similar state-of-the-art CNN models (Dosovitskiy et al., 2020a). The transformer’s self attention like architecture results in it having better generalization properties (Bai et al., 2021). It has been shown that combining CNN embeddings with the vision transformer results in the hybrid system performing better than larger and deeper vision transformer, with less than half the computational fine-tuning cost (Dosovitskiy et al., 2020b). Thus combining the benefits of both architectures we are able to train on lesser data and get improved accuracy.

Since focal length in pixels has an unbounded range and it changes whenever one resizes images, we estimate vertical field of view (vfov)  $v$  in radians and convert it to focal length  $f_y$  via:

$$f_y = \frac{0.5h}{\tan(0.5v)} \quad (6)$$

where  $h$  is the image height in pixels. We follow (Zhu et al., 2020) and (Kocabas et al., 2021c) to assume zero camera yaw and the effective focal length values in both directions are the same, i.e.,  $f_x = f_y = f$ .

TransFocal takes the complete image as input to predict vfov which is the same for all subjects in the image of a video sequence. This means that inference has to only be performed once in order to get absolute coordinates for each frame of a video sequence. The full image contains rich cues that can facilitate transformer self attention. Vanishing points and geometric lines help the network semantically reason the vertical field of view. Similar to our absolute depth map head and (Kocabas et al., 2021c), we discretize the space of vfov  $v$  into  $B$  bins, converting the regression problem into a classification problem. Also similar to our

depth map head, we aggregate the predicted probability mass using a softargmax operation. For the losses, we found in our testing that combining cross entropy loss  $\mathcal{L}_{CE}$  (with a smaller weight) with softargmax-biased-L2 loss (Kocabas et al., 2021c)  $\mathcal{L}_{agmax}$  improves model convergence. The final loss  $\mathcal{L}_{foc}$ :

$$\mathcal{L}_{foc} = \lambda_{agmax}\mathcal{L}_{agmax} + \lambda_{CE}\mathcal{L}_{CE} \quad (7)$$

## 4 IMPLEMENTATION DETAILS

In the following, we will list the implementation details of both the absolute depth map head, Absolute-ROMP, and the focal length estimation network, TransFocal, in terms of network architecture, training setting details, training datasets, and evaluation metrics.

### 4.1 Absolute-ROMP

**Network Architecture.** We employ HRNet-32 (Wang et al., 2019) as the backbone for Absolute-ROMP. We also maintain CoordConv (Liu et al., 2018) 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. This feature set is then used as input to the four heads for complete end to end prediction. The architecture of absolute depth map is similar to the other map heads. The only difference being the output of the final  $1 \times 1$  convolutional layer, with a size of  $120 \times 64 \times 64$ : As stated earlier the binning resolution is set to 120 and the map size is  $64 \times 64$ . For details on the architecture of the map heads please refer to (Sun et al., 2020).

**Setting Details.** The input images are resized to  $512 \times 512$ , by keeping the same aspect ratio and padding with zeros. The size of the backbone feature is  $H_b = W_b = 128$ . The maximum number of detection is  $N = 64$ . The learning rate used is  $5e-5$ . The batch size is set to 26. We adopt the Adam optimizer (Kingma and Ba, 2014) for training, and train the model until performance plateaus on validation set.

**Training Datasets.** The basic training datasets we used in the experiments include three 3D pose datasets: (Human3.6M (Ionescu et al., 2014), MPI-INF-3DHP (Mehta et al., 2016), MuCo-3DHP (Mehta et al., 2016)) and four in-the-wild 2D pose datasets (MS COCO (Lin et al., 2014), MPII (Andriluka et al., 2014), LSP (Johnson and Everingham, 2010), Crowdpose (Li et al., 2018)). We also use the pseudo 3D annotations from (Kolotouros et al., 2019) and the pseudo 3D labels of 2D pose datasets provided by

(Joo et al., 2020). For evaluation on a 3D pose dataset 3DPW (von Marcard et al., 2018), we use the train set for fine tuning.

**Evaluation Benchmarks and Metrics.** We evaluate our model on the Human3.6M (Ionescu et al., 2014), Mupots (Mehta et al., 2017) and 3DPW (von Marcard et al., 2018). To evaluate the 3D pose accuracy, we employ mean per joint position error (MPJPE) (Joo et al., 2015) and Procrustes-aligned MPJPE (PM-PJPE). For root depth estimation we employ  $MRPE_z$  (Moon et al., 2019) and  $PCK_{abs}$  (Moon et al., 2019).

MPJPE is the average Euclidean distance between the location of real-life joints on human bodies and the location of predicted joints on 3D pose after translating the root joint ('pelvis') of estimated bodies to the groundtruth root. Procrustes-aligned MPJPE (PMPJPE) uses procrustes alignment (Luo and Hancock, 1999) (PA) to solve for translation, scale and rotation between the estimated bodies and the ground truth and thus mostly focuses on the pose error. Mean root position error ( $MRPE$ ) is the mean of the euclidean distance between the estimated coordinates of the predicted absolute root and ground truth absolute root. The 3D percentage of correct absolute keypoints ( $3DPCK_{abs}$ ) 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 is shown in Figure 2. Similar to (Yang et al., 2021), we use ResNet (He et al., 2015) 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, which means that the input sequence is obtained by simply flattening the spatial dimensions of the ResNet features and projecting to the Transformer dimension. Since the image is not converted into patches directly, the original physical meaning of positional embeddings is missing. Therefore they are not used when mapping the patches into a latent embedding space  $l$  using a linear projection.

We also show the overview of the transformer encoder architecture in Figure 2, to help with the explanation stated below. The input of the first Transformer layer  $z_0$  is calculated as follow:

$$z_0 = l^1 E; l^2 E; l^3 E \dots; l^n E \quad (8)$$

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

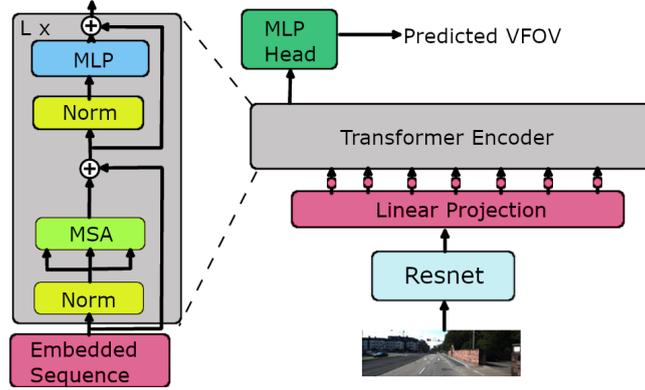


Figure 2: TransFocal Architecture. The image is input into a ResNet backbone to create embeddings which are then projected into latent embedding space and used as input for the vision transformer. The output from the transformer’s many layers is then decoded using a fully connected layer.

specifically the ViT-B16 (Dosovitskiy et al., 2020a) variant. There are  $L$  Transformer layers which consist of multi-headed self-attention (MSA) and multi-layer perceptron (MLP) blocks. At each transformer layer  $\ell$ , the input of the self-attention block is a triplet of  $Q$  (query),  $K$  (key), and  $V$  (value), which 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:

$$AH = \text{softmax}\left(\frac{Q \times K^T}{\sqrt{d}}\right)V \quad (9)$$

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

$$MSA(z_{\ell-1}) = z_{\ell-1} + \text{concat}(AH_1; AH_2; \dots; AH_m) \times W_o, \quad (10)$$

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

$$z_\ell = MLP(\text{Norm}(MSA(z_{\ell-1}))) + MSA(z_{\ell-1}) \quad (11)$$

where  $Norm$  means the layer normalization operator. Finally we use a fully connected layer to decode the output of the vision transformer.

**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 (Kingma and Ba, 2014) for training. We train the model until performance plateaus on validation set.

**Training and Evaluation Datasets.** A dataset is created using the Pano360 dataset (Kocabas et al., 2021c). Pano360 consists of real panoramas taken

from flickr (Flickr, 2022) and 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 for creating the dataset is available at (Kocabas et al., 2021b). Note that as the image generation parameters are randomized, it is difficult to recreate exact the same dataset.

## 5 EXPERIMENTS AND RESULTS

Here we report some performance comparison on multiple datasets and a ablation study in using the Absolute-ROMP.

### 5.1 Performance Comparison

The root joint localization results on Human3.6M dataset are shown in Table 1. The baselines reported in the first 3 rows follow a two-stage approach (Moon et al., 2019), where 2D pose and 3D pose are estimated separately, and 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

Table 1:  $MRPE_z$  results comparison with state-of-the-art on the Human3.6M dataset.

Methods	$MRPE_z$
Baseline	261.9
Baseline w/o limb joints	220.2
Baseline with RANSAC	207.1
RootNet (Moon et al., 2019)	108.1
HDNET (Lin and Lee, 2020)	69.9
Ours	<b>68</b>

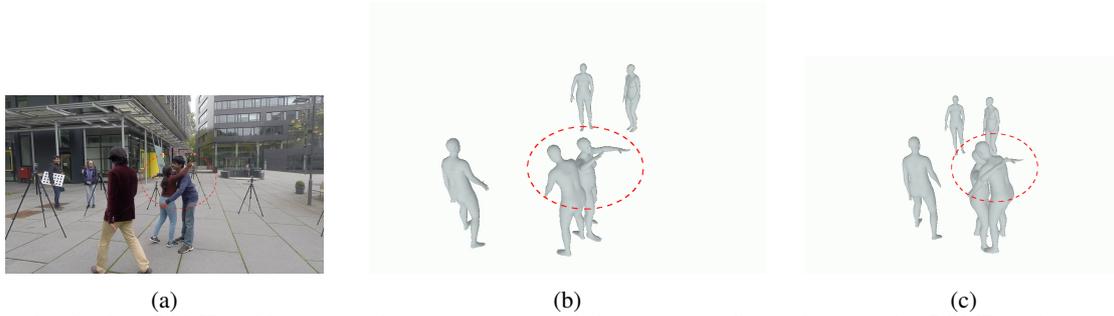


Figure 3: Absolute-ROMP is able to correctly position two people hugging: (a). Original image. (b). ROMP mesh positioning using camera parameters (Sun et al., 2020). (c). Absolute-ROMP mesh positioning using absolute depth prediction.

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 the figures reported in (Moon et al., 2019).

We also compare with the state-of-the-art approaches (Moon et al., 2019; Lin and Lee, 2020). In (Moon et al., 2019) a multi-stage approach is used while in (Lin and Lee, 2020) a graph convolution network model is used. Our model is able to outperform the SOTA while maintaining inference in real time. Our system runs at  $\sim 23$  fps on a Nvidia Tesla V100 GPU. Our root joint localization head achieves a 68 mm accuracy.

We also showcase the performance of Absolute-ROMP on the MUPOTS dataset in Table 2. Our model has comparable performance to the SOTA. We also compare MJPJPE and PMPJPE performance in Table 3 which does not make use of the absolute depth as joints are evaluated after root alignment. Even with added depth map prediction, we are able to maintain comparable performance with the SOTA. We also show qualitative comparison with ROMP in Figure 3 which 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.

Table 2: Results on the MuPoTS-3D dataset.

Methods	$PCK_{abs}$
Moon et al. (Moon et al., 2019)	31.9
HDNET (Lin and Lee, 2020)	35.2
SMAP (Zhen et al., 2020)	<b>35.4</b>
Ours	35.28

Table 3: Comparisons to the state-of-the-art methods on 3DPW.

Methods	MPJPE	PMPJPE
YOLO + VIBE (Kocabas et al., 2019)	82.9	51.9
ROMP(HRNET-32) (Sun et al., 2020)	<b>76.7</b>	<b>47.3</b>
Absolute-ROMP(HRNET-32)	84	50.5

Finally, in Table 4 we show case TransFocal performance against a SOTA approach CamCalib (Kocabas et al., 2021c). Our model is trained on a dataset created from partially available images of the Pano360 dataset, while we compare the performance with a model provided by the author (Kocabas et al., 2021c) pretrained on dataset created from complete Pano360 dataset. We evaluate on a portion of the dataset unseen by our model. As can be seen our model is able to outperform CamCalib by up to 40%.

Table 4: vfov error results comparison with state-of-the-art on dataset created from Pano360 dataset.

Methods	vfov diff(degrees)
CamCalib (Kocabas et al., 2021c)	26.35
TransFocal	<b>15.59</b>

## 5.2 Ablation Study

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 5. 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 .

Table 5: vfov error comparison after 1 epoch with different losses for supervision.

Methods	vfov diff
softargmax-biased-L2	15.8
softargmax-biased-L2+cross entropy	<b>15.6</b>

## 6 CONCLUSION

We build upon an end to end one-stage network ROMP for monocular multi-person 3D mesh regression, by adding in absolute distance prediction to cre-

ate our Absolute-ROMP. In order to eliminate the need for intrinsic parameters of the camera, we also design and train a focal length prediction network called TransFocal, which is a CNN+Transformer hybrid model. We evaluate our models on industry standard benchmarks. Absolute-ROMP shows competitive performance while maintaining real time performance, while TransFocal is able to outperform the state-of-the-art.

Future work could incorporate the absolute camera parameters eliminating 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 (Patel et al., 2021). In addition, real-time clothes and texture prediction would be especially beneficial for VR and AR applications.

## ACKNOWLEDGEMENTS

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 #1737533), 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).

## REFERENCES

- Aitpayev, K. and Gaber, J. (2020). Creation of 3d human avatar using kinect. In *Asian Transactions on Fundamentals of Electronics, Communication & Multimedia*.
- Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. (2014). 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693.
- Bai, Y., Mei, J., Yuille, A., and Xie, C. (2021). Are transformers more robust than CNNs? In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.
- Benzine, A., Chabot, F., Luvison, B., Pham, Q. C., and Achrd, C. (2021). Pandanet : Anchor-based single-shot multi-person 3d pose estimation. arXiv.
- Boulic, R., Bécheiraz, P., Emering, L., and Thalmann, D. (1997). Integration of motion control techniques for virtual human and avatar real-time animation. pages 111–118.
- Carbonera Luvizon, D., Tabia, H., and Picard, D. (2017). Learning features combination for human action recognition from skeleton sequences. *Pattern Recognition Letters*, 99:13–20. User Profiling and Behavior Adaptation for Human-Robot Interaction.
- Choi, H., Moon, G., and Lee, K. M. (2020). Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose.
- Dabral, R., Gundavarapu, N. B., Mitra, R., Sharma, A., Ramakrishnan, G., and Jain, A. (2019). Multi-person 3d human pose estimation from monocular images.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020a). An image is worth 16x16 words: Transformers for image recognition at scale.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020b). vision transformer github repository. *GitHub repository*.
- Du, G. and Zhang, P. (2014). Markerless human-robot interface for dual robot manipulators using kinect sensor. *Robotics and Computer-Integrated Manufacturing*, 30(2):150–159.
- Fang, H.-S., Cao, J., Tai, Y.-W., and Lu, C. (2018). Pairwise body-part attention for recognizing human-object interactions.
- Flickr (2022). Yahoo. <https://www.flickr.com/>.
- Grompone von Gioi, R., Jakubowicz, J., Morel, J.-M., and Randall, G. (2010). Lsd: A fast line segment detector with a false detection control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):722–732.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.
- Hold-Geoffroy, Y., Sunkavalli, K., Eisenmann, J., Fisher, M., Gambaretto, E., Hadap, S., and Lalonde, J.-F. (2017). A perceptual measure for deep single image camera calibration.
- Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014). Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339.
- Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., and Daniilidis, K. (2020). Coherent reconstruction of multiple humans from a single image.
- Johnson, S. and Everingham, M. (2010). Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, pages 12.1–12.11. BMVA Press. doi:10.5244/C.24.12.
- Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., and Sheikh, Y. (2015). Panoptic studio: A massively multiview system for social motion capture. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3334–3342.
- Joo, H., Neverova, N., and Vedaldi, A. (2020). Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation.

- Kanazawa, A., Black, M. J., Jacobs, D. W., and Malik, J. (2017). End-to-end recovery of human shape and pose.
- Kanazawa, A., Zhang, J. Y., Felsen, P., and Malik, J. (2018). Learning 3d human dynamics from video.
- Kendall, A., Grimes, M., and Cipolla, R. (2015). PoseNet: A convolutional network for real-time 6-dof camera re-localization. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2938–2946.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.
- Kocabas, M., Athanasiou, N., and Black, M. J. (2019). Vibe: Video inference for human body pose and shape estimation.
- Kocabas, M., Huang, C.-H. P., Hilliges, O., and Black, M. J. (2021a). Pare: Part attention regressor for 3d human body estimation.
- Kocabas, M., Huang, C.-H. P., Tesch, J., Müller, L., Hilliges, O., and Black, M. J. (2021b). Spec github repository. *GitHub repository*.
- Kocabas, M., Huang, C.-H. P., Tesch, J., Müller, L., Hilliges, O., and Black, M. J. (2021c). Spec: Seeing people in the wild with an estimated camera.
- Kolotouros, N., Pavlakos, G., Black, M. J., and Daniilidis, K. (2019). Learning to reconstruct 3d human pose and shape via model-fitting in the loop.
- Lee, J., Go, H., Lee, H., Cho, S., Sung, M., and Kim, J. (2021). Ctrl-c: Camera calibration transformer with line-classification.
- Lee, J.-H. and Kim, C.-S. (2019). Monocular depth estimation using relative depth maps. pages 9721–9730.
- Li, J., Wang, C., Liu, W., Qian, C., and Lu, C. (2020a). Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation.
- Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.-S., and Lu, C. (2018). Crowdpose: Efficient crowded scenes pose estimation and a new benchmark.
- Li, Y.-L., Liu, X., Lu, H., Wang, S., Liu, J., Li, J., and Lu, C. (2020b). Detailed 2d-3d joint representation for human-object interaction.
- Li, Y.-L., Xu, L., Liu, X., Huang, X., Xu, Y., Wang, S., Fang, H.-S., Ma, Z., Chen, M., and Lu, C. (2020c). Pastanet: Toward human activity knowledge engine.
- Li, Z., Dekel, T., Cole, F., Tucker, R., Snavely, N., Liu, C., and Freeman, W. T. (2019). Learning the depths of moving people by watching frozen people.
- Li, Z. and Snavely, N. (2018). Megadepth: Learning single-view depth prediction from internet photos.
- Lin, J. and Lee, G. H. (2020). Hdnet: Human depth estimation for multi-person camera-space localization.
- Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2014). Microsoft coco: Common objects in context.
- Liu, R., Lehman, J., Molino, P., Such, F. P., Frank, E., Sergeev, A., and Yosinski, J. (2018). An intriguing failing of convolutional neural networks and the co-ordconv solution.
- Lo Presti, L. and La Cascia, M. (2016). 3d skeleton-based human action classification: A survey. *Pattern Recognition*, 53:130–147.
- Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. (2015). Smpl: a skinned multi-person linear model. volume 34.
- Luo, B. and Hancock (1999). Feature matching with procrustes alignment and graph editing. In *Image Processing And Its Applications, 1999. Seventh International Conference on (Conf. Publ. No. 465)*, volume 1, pages 72–76 vol.1.
- Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., and Theobalt, C. (2016). Monocular 3d human pose estimation in the wild using improved cnn supervision.
- Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., and Theobalt, C. (2017). Single-shot multi-person 3d pose estimation from monocular rgb.
- Moon, G., Chang, J. Y., and Lee, K. M. (2019). Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image.
- Patel, P., Huang, C.-H. P., Tesch, J., Hoffmann, D. T., Tripathi, S., and Black, M. J. (2021). AGORA: Avatars in geography optimized for regression analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Qi, S., Wang, W., Jia, B., Shen, J., and Zhu, S.-C. (2018). Learning human-object interactions by graph parsing neural networks.
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks.
- Rogez, G., Weinzaepfel, P., and Schmid, C. (2019). LCR-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.
- Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M. J., and Mei, T. (2020). Monocular, one-stage, regression of multiple 3d people.
- von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., and Pons-Moll, G. (2018). Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*.
- Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., and Xiao, B. (2019). Deep high-resolution representation learning for visual recognition.
- Workman, S., Greenwell, C., Zhai, M., Baltenberger, R., and Jacobs, N. (2015). Deepfocal: A method for direct focal length estimation. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 1369–1373.
- Workman, S., Zhai, M., and Jacobs, N. (2016). Horizon lines in the wild.
- Xu, Y., Zhu, S.-C., and Tung, T. (2019). Denserac: Joint 3d pose and shape estimation by dense render-and-compare.

- Yang, G., Tang, H., Ding, M., Sebe, N., and Ricci, E. (2021). Transformer-based attention networks for continuous pixel-wise prediction.
- Zanfir, A., Marinoiu, E., and Sminchisescu, C. (2018a). Monocular 3d pose and shape estimation of multiple people in natural scenes: The importance of multiple scene constraints. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2148–2157.
- Zanfir, A., Marinoiu, E., Zanfir, M., Popa, A.-I., and Sminchisescu, C. (2018b). 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., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Zhen, J., Fang, Q., Sun, J., Liu, W., Jiang, W., Bao, H., and Zhou, X. (2020). Smap: Single-shot multi-person absolute 3d pose estimation.
- Zhu, R., Yang, X., Hold-Geoffroy, Y., Perazzi, F., Eisenmann, J., Sunkavalli, K., and Chandraker, M. (2020). Single view metrology in the wild.
- Zimmermann, C., Welschehold, T., Dornhege, C., Burgard, W., and Brox, T. (2018). 3d human pose estimation in rgb-d images for robotic task learning.

