D-PoSE: Depth as an Intermediate Representation for 3D Human Pose and Shape Estimation

Nikolaos Vasilikopoulos
Computer Science Department
University of Crete
and
Institute of Computer Science, FORTH
Heraklion, Greece
vassilicopoulos@gmail.com

Drosakis Drosakis
Institute of Computer Science, FORTH
Heraklion, Greece
drosakis@ics.forth.gr

Antonis Argyros
Computer Science Department
University of Crete
and
Institute of Computer Science, FORTH
Heraklion, Greece
argyros@ics.forth.gr

***Abstract*—Markerless 3D capture of human pose and shape is a key enabler for immersive Virtual, Augmented, and Extended Reality (VR/AR/XR) experiences such as avatar embodiment, telepresence, and motion-driven interaction, where capture should ideally rely on the commodity RGB cameras already present in phones and headsets rather than on depth sensors or multi-camera rigs. We present D-PoSE (Depth as an Intermediate Representation for 3D Human Pose and Shape Estimation), a one-stage method that estimates human pose and SMPL-X shape parameters from a single RGB image. Recent works use larger models with transformer backbones and decoders to improve the accuracy in human pose and shape (HPS) benchmarks. D-PoSE proposes a vision-based approach that uses the estimated human depth-maps as an intermediate representation for HPS and leverages training with synthetic data and the ground-truth depth-maps provided with them for depth supervision during training. Although trained on synthetic datasets, D-PoSE achieves state-of-the-art performance on the real-world benchmark datasets, EMDB and 3DPW. Despite its simple lightweight design and the CNN backbone, it outperforms ViT-based models that have a number of parameters that is larger by almost an order of magnitude. Its lightweight design and single-image input make it well suited to interactive VR/AR/XR systems operating under limited compute budgets.**

I. Introduction

Vision-based 3D human pose and shape (3D HPS) estimation is a key enabling technology for immersive Virtual, Augmented, and Extended Reality (VR/AR/XR) applications. Avatar animation, telepresence, embodied interaction, and motion analysis, all require recovering the 3D configuration and body shape of a person, ideally from the commodity RGB cameras already present in phones and headsets, rather than from specialised depth sensors or multi-camera capture rigs. While there is already a number of effective solutions for the problem of 2D human body joints estimation from RGB images that are based on neural network architectures [1]–[3], the emphasis has moved to the more demanding problems of 3D pose [4], [5] and 3D mesh estimation [6]–[10] for the whole body and its parts [11], [12], which provide the full geometry needed to render and animate believable virtual humans.

This work was co-funded by the European Union (EUHE MAGICIAN - Grant Agreement 101120731).

Fig. 1. D-PoSE, the proposed 3D human Pose and Shape Estimation method, receives a single RGB image as input (top), produces intermediate depth and part segmentation representations (middle-left and middle-right, respectively) so as to deliver the 3D pose and shape of the imaged person (bottom, side-view of the mesh for visualization purposes). Despite entailing a small fraction of the parameters of current models, D-PoSE outperforms the current state of the art in 3D pose and shape estimation accuracy in the major relevant datasets (3DPW, EMDB).

Still, despite the plethora of approaches that have already been proposed, 3D HPS estimation remains a challenging task. Several approaches use video as input [13], [14], or depth information provided by RGB-D cameras [15]. The recovery of human 3D pose and shape from a single RGB image lacks temporal and depth information and, thus, has to rely on minimal information to solve very challenging 2D to 3D ambiguities. Therefore, this is the most challenging of all settings. At the same time, it is the most general setting and makes the least amount of assumptions regarding the input of the estimation problem: it requires neither the temporal context of video nor the calibrated depth of an RGB-D or multi-camera setup. As a result, a robust and accurate solution given the minimal input of a single image frame can be very impactful for VR/AR/XR systems that target widely available consumer hardware.

In this work, we focus on this challenging version of the 3D HPS estimation problem where 3D human pose and shape have to be recovered on the basis of a single RGB frame, only (see Fig. 1). To address this problem, we propose D-PoSE, a method that leverages ground-truth depth maps from recent synthetic datasets and learns to predict human depth maps that incorporates them in the prediction procedure for more accurate 3D HPS estimation. Specifically, D-PoSE uses synthetic RGB data as input, together with the associated depth maps *which are only used for supervision during training and not as input at run-time.* Human depth maps serve as an intermediate representation, together with an estimated human body parts segmentation.

The training of D-PoSE capitalizes on the availability of synthetic data. With the introduction of the recent BEDLAM synthetic dataset [12], models are able to train only with synthetic data and outperform the accuracy of training with real-world data. BEDLAM provides accurate synthetic depth maps with ground-truth 3D keypoints and SMPL-X [16] parameters. Although there is a domain gap between synthetic and real-world depth, the proposed model generalizes well in real-world datasets.

Current state-of-the-art methods [17], [18] use ViT [19] backbones. While those backbones benefit from large datasets, they increase dramatically the model size. Therefore, those methods need long training times and multiple flagship GPUs, and their size and latency hinder deployment in interactive, resource-constrained VR/AR/XR settings. One of the goals of our work is to provide a lightweight vision-based solution to the 3D HPS estimation problem that has state-of-the-art performance without the need of extra training dataset(s) and does not employ oversized models, so that accurate human capture becomes practical on the kind of hardware available to end users.

We demonstrate that the use of depth information as an intermediate representation together with part segmentation on a simple CNN backbone suffices to deliver state of the art results in terms of both accuracy and model size. Specifically, we performed several experiments on the challenging 3DPW [20] and EMDB [21] datasets. The experimental results demonstrate improvements of 3.0mm in PA-MPJPE, 3.1mm in MPJPE and 3.6mm in MVE when compared with BEDLAM-CLIFF [12] in the challenging 3DPW dataset. When compared with the state-of-the-art method TokenHMR [17] which employs a ViT backbone, our method reduces error by 0.4mm in PA-MPJPE, 2.7mm in MPJPE and 4.3mm in MVE. At the same time, our model has 83.8% less parameters than TokenHMR. Thus, compared to the current state of the art, D-PoSE is both more accurate and more light-weight.

In summary, the main contributions of this work are the following:

- We propose D-PoSE, a novel method to the problem of 3D HPS estimation from a single RGB frame. D-PoSE uses depth information from synthetic data as an intermediate representation and generalizes well to real-world data.
- We demonstrate that D-PoSE achieves state-of-the-art accuracy in Mean Vertices Error (MVE) and Mean per Joint Position Error (MPJPE) in standard HPS benchmarks.
- We also demonstrate that D-PoSE entails significantly less trainable model parameters, specifically 83.8% less parameters compared to the current state-of-the-art method.

II. Related Work

The 3D HPS estimation problem has been approached in several ways, including optimization-based techniques (usually by fitting a mesh to 2D keypoints), learning-based techniques (where a model is trained to predict a 3D mesh). We also review various intermediate representations that have been employed as well as training datasets that are relevant to our work. D-PoSE is a one-stage, learning-based method which takes a single RGB image as input and uses intermediate representations before estimating the 3D mesh.

Optimization-based methods: Optimization approaches use 2D image cues to fit a parametric model. Bogo [22] proposed Simplify, which optimizes the 3D shape and pose of the SMPL [23] human model using 2D keypoints. Omran [24] proposed the use of silhouettes to handle perspective ambiguities. Lassner [25] used part segmentation to improve the body shape and pose estimation. Optimization approaches require less data but are prone to 2D-3D ambiguities.

Learning-based methods: Learning based approaches estimate directly model parameters [9], [10], [18], [26]–[30]. A model-free representation can be estimated such as vertices [31]–[33] or implicit shape [34]–[36]. Li [37] proposed a novel hybrid inverse kinematics solution (HybrIK) which computes the 3D joint positions of a human body by combining an analytical solution and a neural network regression. Pose priors can also be employed, imposing constraints on the human pose and shape in order to reduce invalid estimations. These could include joint limits [38] where they would prune invalid human poses, Gaussian Mixture Models [22], Generative Adversarial Networks [18], [39], VAEs [40] and normalizing flows [41] that can be used as knowledge priors in the training process. Kolotouros [8] proposed SPIN which improves the pose estimation accuracy by fitting the body model to 2D keypoints in the training loop. CLIFF [9] provides the neural network with information about the bounding box coordinates containing the human in the image, gaining a noticeable accuracy improvement. These methods encounter less ambiguities but rely on additional data for robust training.

Intermediate representations: Intermediate representations could allow for more training data to be injected in the training process. HoloPose [42] aligns initial 3D part based model prediction with the 2D keypoints, 3D keypoints and DensePose [6]. More recently, Kocabas [10] proposed PARE, a model that uses a part-segmentation branch together with an attention mechanism in order to achieve an occlusion-robust method. Our part-segmentation branch is highly inspired by PARE but deviates considerably from it with respect to specific choices in its architecture. Zhu [43], [44] (HMD) argued

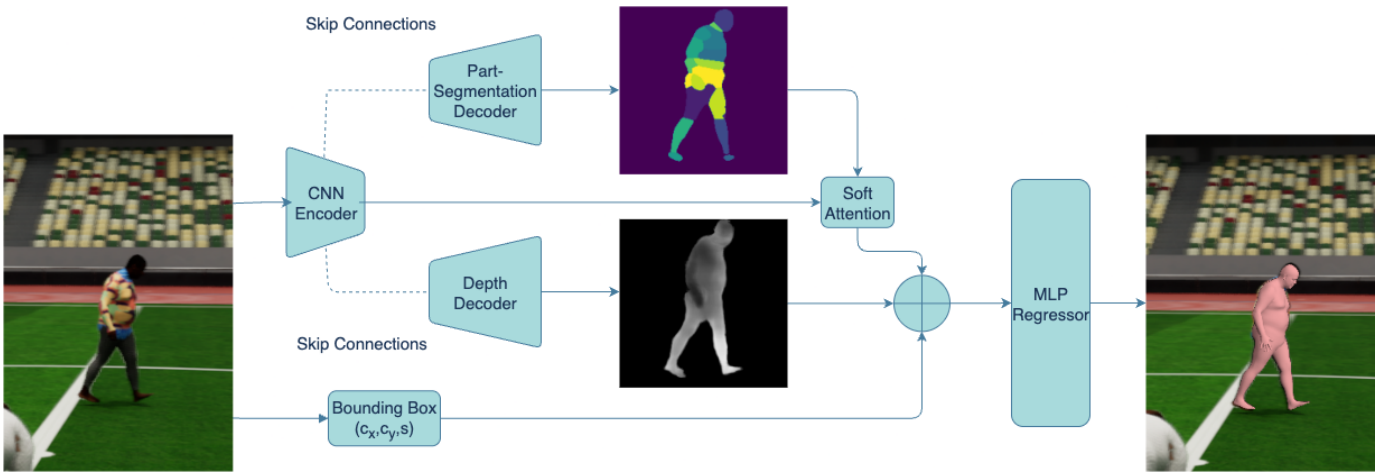

Fig. 2. The architecture of D-PoSE. Given an input image, features are extracted using a CNN. With these feature maps a human depth map and a part-segmentation map are estimated. The original features pass through a soft-attention mechanism which uses part-segmentation maps. The final features are concatenated with the bounding-box information and the depth features and are given as input to the regressor which estimates the 3D human pose and shape.

that by utilizing per-pixel shading information and depth, it is possible to refine the shape and produce a detailed 3D mesh with deformations. Although HMD suggests the use of depth, our method directly uses it in pose and shape estimation process and does not deform the SMPL-X mesh based on the depth.

Depth estimation: Varol [45] suggested to train a CNN with synthetic data and use them to to predict human depth maps and human part-segmentation maps. However neither the human depth nor the segmentation map was used for pose or shape estimation. Zhou found that DIFFNet [46] with HRNet as encoder and a UNET-like depth decoder is effective in standard depth estimation datasets.

Synthetic data: AGORA [47] provides SMPL-X ground truth data and synthetic images of clothed humans generated from static commercial scans. The inclusion of AGORA in training datasets enhances the accuracy of 3D HPS estimation methods. Additionally, AGORA serves as a benchmark for evaluating 3D HPS estimation approaches.

Black [12] proposed BEDLAM, a new synthetic dataset with ground truth SMPL-X data, realistic human images and depth maps. Training HMR [18] and CLIFF [9] on the BEDLAM dataset proves itself enough to achieve state-of-the-art performance. The same work also suggests the use of vertices loss.

Vision transformers backbone: HMR2.0 [48] uses ViT backbone to encode the image and a transformer based decoder to predict the 3D mesh. TokenHMR [17], using the same backbone with HMR2.0, reformulates the problem of HPS by tokenizing the pose in the encoder and letting the decoder reconstruct the original pose.

The proposed D-PoSE approach: D-PoSE is a one-stage method that takes a single RGB image as input and estimates two intermediate representations: (1) human depth and (2) part-segmentation of the human. Using these representations, along with the original CNN features, it regresses the 3D human pose and shape. D-PoSE does not use vision transformers as backbone but a CNN.

III. Methodology

An overview of the architecture of D-PoSE is provided in Fig. 2. Given an input image, features are extracted using a CNN. With these feature maps a human depth map and a part-segmentation map are estimated. The CNN features pass through a soft-attention mechanism which uses the part-segmentation maps. The final features are concatenated with the bounding-box information and the estimated human depth map and are given as input to the regressor which estimates the 3D human pose and shape. Below, we provide further details on each and every of the aforementioned modules and representations.

A. CNN Encoder

D-PoSE uses the High-Resolution Network (HRNet-W48) [49]–[51] as the convolutional neural network (CNN) encoder. The HRNet-W48 is selected for its ability to produce spatially precise feature maps at multiple resolutions. Specifically, given a single RGB image input of dimensions 224×224, the encoder generates a set of feature maps at four distinct resolutions: $\mathbf{F}_1 \in \mathbb{R}^{384\times7\times7}$, $\mathbf{F}_2 \in \mathbb{R}^{192\times14\times14}$, $\mathbf{F}_3 \in \mathbb{R}^{96\times28\times28}$ and $\mathbf{F}_4 \in \mathbb{R}^{48\times56\times56}$. These feature maps are utilized in skip-connections with the decoder layers to enhance the reconstruction process. Additionally, the encoder outputs an up-sampled feature vector $\mathbf{F}_{\text{down}} \in \mathbb{R}^{720\times56\times56}$. HRNet-W48 has previously been used in BEDLAM-CLIFF [12] and BEDLAM-HMR [12] and is a standard choice for CNN backbones used in pose estimation tasks.

B. Human Models

To predict the 3D human mesh, we utilize the SMPL-X [16] body model, which consists of $N = 10{,}475$ vertices and $K = 54$ joints, including those for the neck, jaw, eyeballs and fingers. The SMPL-X model is represented by the function $M(\theta, \beta, \psi)$,

where θ denotes pose parameters, β captures shape parameters, and ψ represents facial expression parameters.

For evaluations on the 3DPW [20] and EMDB [21] datasets, we transform our predicted SMPL-X [16] meshes into SMPL [23] format by applying a vertex mapping matrix $D \in \mathbb{R}^{10475\times6890}$. This conversion is used exclusively for assessing body pose and shape. Similarly, we convert the ground truth SMPL-X vertices to SMPL format using D after neutralizing the hand and face poses. To calculate joint errors, we extract 22 joints from the vertices using the SMPL joint regressor.

For both SMPL and SMPL-X we use the gender neutral models.

C. Loss Function

The loss function L consists of three parts, depth loss, segmentation loss and 3D human loss:

$$L = L_{depth} + L_{segm} + L_{human}. \tag{1}$$

Depth loss: For the depth term loss, background is ignored and a combination of L1 loss and structural similarity index measure is used:

$$L_{depth} = \lambda_1 |depth_{gt} - depth_{pred}| + \lambda_2 (1 - SSIM(depth_{gt}, depth_{pred})). \tag{2}$$

Part segmentation loss: To produce accurate part-segmentation, we use cross entropy loss between the predicted and ground-truth SMPL part-segmentations:

$$L_{segm} = \lambda_3 CrossEntropy(gt, pred). \tag{3}$$

3D human loss: For 3D human prediction, as proposed by BEDLAM-CLIFF [12], we use two MSE SMPL-X losses one for SMPL-X pose θ parameters and one for SMPL-X shape β parameters, a 3D joints MSE loss between the ground-truth and estimated 3D joints, and the newly proposed 3D vertices loss which is a L1 loss between the ground-truth SMPL-X vertices and the estimated SMPL-X vertices:

$$L_{SMPL_{pose}} = ||\hat{\theta} - \theta||,$$

$$L_{SMPL_{shape}} = ||\hat{\beta} - \beta||,$$

$$L_{J3D} = ||\hat{J}_{3D} - J_{3D}||,$$

$$L_{J2D} = ||\hat{J}_{2D} - J_{2D}||,$$

$$L_{V3D} = |\hat{V} - V|.$$

Given the above, the 3D human loss is defined as:

$$L_{human} = \lambda_4 L_{SMPLpose} + \lambda_5 L_{SMPLshape} + \lambda_6 L_{J3D} + \lambda_7 L_{V3D} + \lambda_8 L_{J2D}. \tag{4}$$

D. Depth

Depth is estimated by the depth decoder. The depth decoder takes as input the features extracted by the HRNet backbone and uses skip connections from the previous stages to capture hierarchical features [$\mathbf{F}_0 \in \mathbb{R}^{384\times7\times7}$, $\mathbf{F}_1 \in \mathbb{R}^{192\times14\times14}$, $\mathbf{F}_2 \in \mathbb{R}^{96\times28\times28}$ and $\mathbf{F}_3 \in \mathbb{R}^{48\times56\times56}$]. The depth decoder along with the CNN encoder have a U-Net [52] like structure. However, the decoder is lighter and not symmetrical to the encoder path. The output of the decoder is the relative depth of the human ignoring the background which is represented with zero values. The BEDLAM dataset provides ground truth depth maps stored as 32-bit float in Unreal coordinate system units. From the depth maps we remove the background using the background mask provided and keep only the values of the human body. Then we set background values to zero and normalize the rest of the values in the range $[0.1, 1.0]$.

E. Part Segmentation

Part segmentation maps are estimated by the part-segmentation decoder which is similar to the depth decoder. The difference with the depth-decoder is the last layer which outputs 23 channels and the body model. PARE [10] uses a simple CNN decoder as opposed to ours which uses skip-connections from the encoder forming a UNet like structure. While PARE uses SMPL model, we use SMPL-X since we train with the AGORA and BEDLAM datasets. In contrast to PARE, part-segmentation remains supervised throughout the entire training process and is not disabled at any point.

Since BEDLAM and AGORA provide ground-truth SMPL-X parameters, during training we use these parameters to generate SMPL-X mesh. From the generated ground-truth mesh, we map each vertex to the human joint that it belongs. We end up with 22 different body parts (PARE that uses SMPL has 24), each assigned with a different value (see Fig. 3). The background is also assigned to the value of zero. Finally, we render the part-segmented SMPL-X mesh and use it to supervise the part-segmentation.

F. Soft-Attention

The soft-attention mechanism employed in our work is similar to that used in PARE. It takes as input a tensor $\mathbf{F}_{\text{upsampled}} \in \mathbb{R}^{720\times56\times56}$ which is the CNN features and the estimated the part-segmentation images $\mathbf{S} \in \mathbb{R}^{23\times56\times56}$. The part-segmentation tensor $\mathbf{S}$ is passed through a softmax operation over the spatial dimensions while ignoring the first segmentation-map which is attributed to background, producing normalized attention maps $\sigma(\mathbf{S}) \in \mathbb{R}^{22\times(56\times56)}$.

The attention-weighted features are obtained by:

$$A = \sigma(S) \cdot F^T_{\text{reshaped}}. \tag{5}$$

For the shape of tensor A it holds that $A \in \mathbb{R}^{22\times720}$.

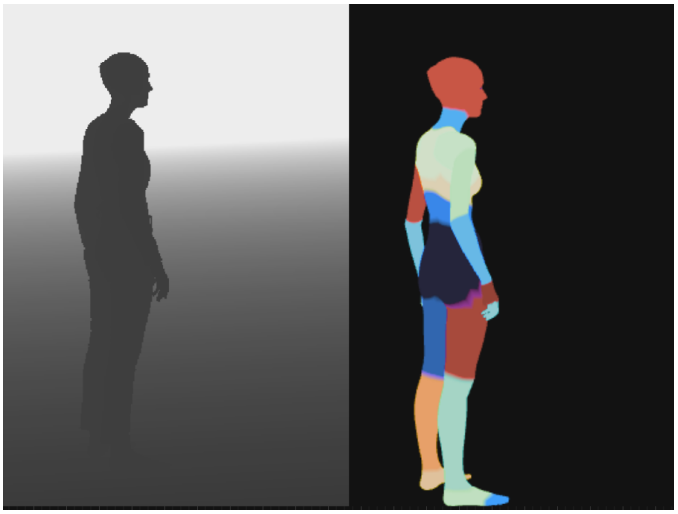

Fig. 3. Left: Ground-truth depth-pap visualized in grayscale (BEDLAM dataset). Right: Ground-Truth SMPL-X Mesh after rendering with part-segmentation (BEDLAM dataset).

G. Bounding Box

As proposed by CLIFF [9], we supervise the 2D reprojection loss in the original full-frame image instead of the cropped image. Specifically,

$$J_{2D}^{full} = \Pi J_{3D}^{full} = \Pi(J_{3D} + \mathbf{t}^{full}), \tag{6}$$

where $\mathbf{t}^{full}$ represents the translation relative to the optical center of the original image. Also, we concatenate the bounding-box center and scale with the features produced by the attention mechanism. As a result, the estimated global orientation is improved.

H. Decoder Architecture

Our decoder architecture consists of a sequence of upsampling and refinement modules designed to progressively enhance feature maps extracted from the backbone network. These modules generate either a depth map or a part-segmentation map as the final output.

Input Feature Maps: Let $\mathbf{F}_i \in \mathbb{R}^{B\times C_i\times H_i\times W_i}$ represent the feature maps from the backbone at resolution level i, where B is the batch size, C_i is the number of channels, and $H_i \times W_i$ are the spatial dimensions. We utilize feature maps $\{\mathbf{F}_0, \mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3\}$ across four resolution levels.

Upsampling Modules: Each upsampling module U_i upsamples feature maps from resolution level $i+1$ to level i using bilinear interpolation (scale factor 2) followed by a 1×1 convolution with ReLU activation:

$$\mathbf{F}_i^{\text{up}} = U_i(\mathbf{F}_{i+1}^{\text{up}}) = \text{ReLU}\left(\text{Conv}_{1\times 1}\left(\mathbf{F}_{i+1}^{\text{up}}\right)\right).$$

The upsampled feature map $\mathbf{F}_i^{\text{up}}$ is then concatenated with the corresponding backbone feature map $\mathbf{F}_i$:

$$\mathbf{F}_i^{\text{cat}} = \text{Concat}\left(\mathbf{F}_i^{\text{up}}, \mathbf{F}_i\right).$$

Fusion and Refinement: The concatenated feature maps $\mathbf{F}_i^{\text{cat}}$ are processed by a refinement module R_i, which consists of four residual blocks:

$$\mathbf{F}_i^{\text{ref}} = R_i(\mathbf{F}_i^{\text{cat}}).$$

Final Output Generation: The final output, either a depth map or a part-segmentation map, is produced by applying a series of convolutional layers and ReLU activations to the refined feature map at the lowest resolution level:

$$\mathbf{O} = \text{Conv}_{1\times 1}\left(\text{ReLU}\left(\text{Conv}_{3\times 3}(\mathbf{F}_0^{\text{ref}})\right)\right).$$

Here, $\mathbf{O} \in \mathbb{R}^{B\times C_{\text{out}}\times H_0\times W_0}$, where $C_{\text{out}} = 1$ for depth maps and $C_{\text{out}} = 23$ for part-segmentation maps.

I. Regressor

To regress the SMPL-X pose parameters we use a regressor with the same MultiLinear layer that ReFit [53] proposes. The forward pass is efficiently computed using Einstein summation notation, and bias terms are added per head. The 22 heads representing each joint compute the body pose parameters in parallel.

The three camera parameters and the eleven shape parameters are computed by simple linear layers followed by ReLU activation functions.

Our model demonstrates faster convergence and training speeds with this regressor compared to the PARE [10] and CLIFF [9] regressors.

IV. Experiments

A. Datasets

D-PoSE is trained solely on synthetic data. BEDLAM [12] is used subsampled at 6 frames per second as the proposed method BEDLAM-CLIFF. We use the ground truth training data, including the provided depth maps. Also, BEDLAM provides masks for the background which we use to remove it from the depth maps.

AGORA [47] is the the second synthetic dataset we use for training. AGORA is used to supervise the segmentation and 3D human loss but not the depth since it lacks ground truth depth maps.

Our method is evaluated on the 3DPW [20] and EMDB [21] datasets. Both of them contain real images of humans in the wild. Since both datasets have ground truth SMPL data, we are able to calculate Mean Vertices Error (MVE) on both datasets to capture the accuracy of the estimated human shapes.

We also use the RICH [54] dataset for obtaining qualitative results and for the ablation study (see Table III). RICH differs from the other datasets by including humans interacting with objects and their environment in both indoor and outdoor scenes.

	Training Datasets	Method	EMDB [21]			3DPW [20]		
			MVE	MPJPE	PA-MPJPE	MVE	MPJPE	PA-MPJPE
HRNet	SD	PARE	-	-	-	97.9	82.0	50.9
	SD	CLIFF	-	-	-	87.6	73.9	46.4
	BL	BEDLAM-HMR	-	-	-	93.1	79.0	46.4
	BL	BEDLAM-CLIFF	113.2	97.1	61.3	85.0	72.0	46.6
	BL	**D-PoSE (Ours)**	**99.0**	**85.5**	**53.2**	**81.4**	**68.9**	**43.6**
ViT	BL	HMR2.0	106.6	90.7	51.3	88.4	72.2	45.1
	BL	TokenHMR	106.2	89.6	**49.8**	85.7	71.6	44.0
HR Net	BL	**D-PoSE (Ours)**	**99.0**	**85.5**	53.2	**81.4**	**68.9**	**43.6**

TABLE I

COMPARISON OF HPS ERRORS ON THE EMDB AND 3DPW DATASETS. SD DENOTES STANDARD REALISTIC DATASETS, WHILE BL DENOTES TRAINING EXCLUSIVELY WITH SYNTHETIC DATASETS (BEDLAM AND AGORA). SEE TEXT FOR DETAILS.

Method	GFLOPs	# Params	Infer. time
TokenHMR	254.31	681.0 M	441 ms
D-PoSE (ours)	**64.70**	**81.2 M**	**265 ms**

TABLE II

COMPARING COMPLEXITY OF TOKENHMR AND D-POSE WITH RESPECT TO GFLOPS, NUMBER OF PARAMETERS AND INFERENCE TIME (CPU).

B. Training

We train our model using PyTorch in one stage. Training requires $200K$ iterations with batch size of 64. The optimizer used is Adam with a learning rate of $1e-5$ and zero weight decay. For numerical stability we use gradient clipping with value 1.5. A single NVidia RTX-A100 GPU is used for all the experiments. Training in our system requires 3 days. Random augmentations are applied to the RGB images using Albumentations [55] similarly with BEDLAM-CLIFF. Those augmentations include random cropping, down-scaling, compressing the image, random rain and snow noise, multiplicative noise, motion blur, blurring, random occlusions, CLAHE and equalization, random changes to brightness and contrast, hue saturation, random gamma and posterization.

We use HRNet-W48 as the CNN backbone to extract features from the RGB image in four resolutions. The size of the input RGB image is 224×224. HRNet-W48 is initialized with weights pretrained on COCO [56]. The Neural 3D Mesh Renderer [57] is used to render the part-segmented SMPL mesh during training.

For fair comparison with BEDLAM-CLIFF we use the 80% of BEDLAM and AGORA for training.

The coefficients of the loss functions for the experiments are: $\lambda_1 = 0.1$, $\lambda_2 = 0.02$, $\lambda_3 = 0.1$, $\lambda_4 = 10$, $\lambda_5 = 0.01$, $\lambda_6 = 50$, $\lambda_7 = 10$ and $\lambda_8 = 50$.

C. Evaluation Metrics

For the quantitative evaluation of D-PoSE we use the following well-established evaluation metrics:

Mean Per Joint Position Error (MPJPE): MPJPE aligns the predicted and ground-truth 3D joints at the pelvis and measures the resulting distances, providing a comprehensive evaluation of pose and shape, including global rotations.

Procrustes-Aligned MPJPE (PA-MPJPE): PA-MPJPE applies Procrustes alignment before calculating MPJPE, focusing on articulated pose accuracy by removing scale and rotation discrepancies.

Mean Vertex Error (MVE): MVE also considers pelvis alignment of the predicted and ground-truth 3D joints but evaluates the distances between vertices on the human mesh surface.

D. Quantitative Results

In Table I we compare our method to the current state of the art methods. In order to evaluate our method we convert 3DPW and EMDB SMPL meshed to SMPL-X. In both 3DPW and EMDB we report Mean Vertex Error (MVE) using the vertices obtained from the SMPL mesh, Mean Per Joint Position Error (MPJPE) of the human 3D joints, and Procrustes-Aligned Mean Per Joint Position Error (PA-MPJPE) between the predictions and the ground-truth. All metrics are reported in *mm*.

The results in Table I show that our model in 3DPW reduces PA-MPJPE by 3.0mm, MPJPE by 3.1mm and MVE by 3.6mm when compared with BEDLAM-CLIFF (HRNet backbone). In EMDB reduces PA-MPJPE by 8.1mm, MPJPE by 11.6mm and MVE by 14.2mm when compared with BEDLAM-CLIFF (HRNet backbone).

When compared with TokenHMR (ViT backbone) in 3DPW reduces PA-MPJPE by 0.4mm, MPJPE by 2.7mm and MVE by 4.3mm. In EMDB reduces MPJPE by 4.1mm and MVE by 7.2mm.

Furthermore, the results in Table I demonstrate that training exclusively on synthetic data is effective and generalizes well to real-world data.

In Table II we present the floating point operations, the number of model parameters and the inference time of D-PoSE. All reported figures suggest that compared to the state-of-the-art model TokenHMR, D-PoSE is lighter by a large margin, having 83.8% less parameters. The reason that our

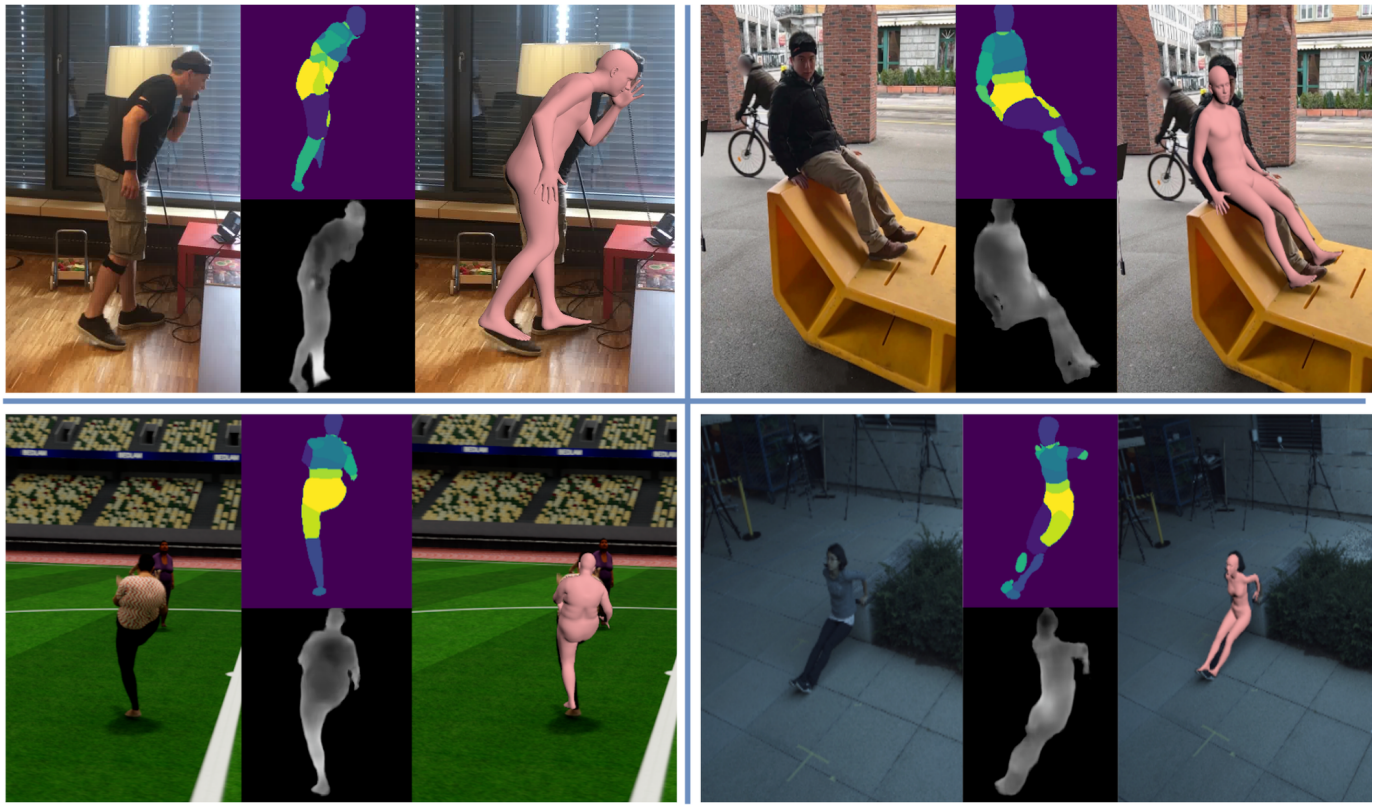

Fig. 4. Each image block represents: the input image (left); the part-segmentation estimation as an intermediate representation (middle-top); the human depth map as an intermediate representation (middle-bottom); the 3D HPS estimation of our method (right). The figure illustrates results from the 3DPW dataset (top left block) the EMDB test set (top right), synthetic image sampled from the BEDLAM validation set (bottom left) and from the RICH dataset (bottom right).

Dataset, Method	PA-MPJPE	MPJPE	MVE
3DPW, w/o Depth	44.3	**68.8**	**81.3**
3DPW, with Depth	**43.6**	68.9	81.4
RICH, w/o Depth	50.1	80.6	92.1
RICH, with Depth	**47.8**	**77.0**	**87.8**
EMDB, w/o Depth	53.5	87.6	101.8
EMDB, with Depth	**53.2**	**85.5**	**99.0**

TABLE III
ABLATION STUDY ON THE IMPACT OF USING (OR NOT) DEPTH IN D-POSE. RESULTS OBTAINED ON THE 3DPW, RICH AND EMDB DATASETS.

model is significantly smaller is that we use a CNN backbone instead of ViT while also using lightweight decoders and regressor.

E. Qualitative Results

Our qualitative results provide evidence on the effectiveness of our method across a diverse set of challenging scenarios. Figure 4 consolidates results from four key datasets, illustrating the versatility and robustness of our approach across a variety of environments and challenges.

The top-left section of Fig. 4 showcases the 3D HPS estimation capabilities of D-PoSE on the 3DPW dataset, along with intermediate representations of depth and part segmentation. Despite the challenges posed by realistic outdoor scenes and occlusions, our method exhibits strong generalization, effectively transferring from synthetic training data to real-world environments. Its robustness is further evidenced by maintaining accuracy even in heavily occluded scenes, a common issue in real-world human pose estimation (HPS) applications. Figure 4 top-right presents results from the EMDB dataset, highlighting our method's performance in a scene with a challenging pose. In the bottom-left of Fig. 4, results from the synthetic BEDLAM dataset illustrate our method's ability to maintain high accuracy, validating its efficacy across both real and synthetic environments. Finally, the bottom-right of Fig. 4 presents results from the RICH dataset, which features complex human poses.

Figure 5 shows D-PoSE in the BEDLAM validation set and 3DPW test set with multiple people in the same image. D-PoSE is able to estimate multiple people in a single pass by leveraging batching.

Qualitative results also showcase the robustness of our method in diverse inputs regarding the race, gender and body-type of the person. These qualitative results underscore the generalization capability of our method and its potential to handle demanding real-world scenarios.

Fig. 5. D-Pose can estimate the 3D shape and pose of multiple people in the same image. The examples are from the BEDLAM validation set (top) and from the 3DPW test set (bottom).

V. Ablation Study

We conduct an ablation study on the impact of using depth in our model architecture. As shown in Table III, the introduction of depth as an intermediate representation improves PA-MPJPE in 3DPW by 0.7mm. In the RICH dataset, MPJPE is reduced by 3.6mm, MVE by 4.3mm and PA-MPJPE by 2.3mm. In the EMDB dataset, MPJPE is reduced by 2.1mm, MVE by 2.8mm and PA-MPJPE by 0.3mm. We consider this a significant improvement in the challenging 3DPW, RICH and EMDB datasets.

In Table IV, we show the scalability of our model when trained on a mixture of synthetic and real data. Although real data lacks ground-truth depth for supervision, D-PoSE with depth consistently outperforms D-PoSE without depth. For this experiment, we used 3DPW, BEDLAM, and AGORA as the training datasets. The evaluation metrics on EMDB further underscore the significant impact of depth in achieving higher 3D accuracy, even in the absence of ground-truth depth labels.

Method	PA-MPJPE	MPJPE	MVE
D-PoSE w/o depth	51.7	87.2	100.7
D-PoSE with depth	**51.0**	**84.7**	**97.4**

TABLE IV
Ablation study on the impact of using (or not) depth when training with both real and synthetic data. Testing is performed of the EMDB dataset [21].

VI. Conclusions

We proposed D-PoSE, a novel architecture for 3D human pose and shape estimation based on a single RGB frame. D-PoSE leverages depth as an intermediate representation, achieving state-of-the-art performance across all error metrics on the challenging 3DPW and EMDB datasets. Despite estimating both part-segmentation and depth maps, our approach significantly reduces the number of parameters compared to state of the art methods. Therefore, overall, D-PoSE is both more accurate and lighter, compared to existing methods. D-PoSE trains in one stage, ensuring a lightweight design that makes it a strong foundation for future advancements in human pose estimation. Its reliance on a single RGB image, together with its small parameter count and low inference cost, also makes D-PoSE a practical building block for markerless avatar embodiment, telepresence, and motion-driven interaction in VR/AR/XR systems that target commodity and on-device hardware. Future work could leverage temporal information from video input and/or incorporate larger transformer-based backbones such as Vision Transformers (ViT), and integrate D-PoSE into real-time XR avatar pipelines.

References

[1] Z. e. a. Cao, "Realtime multi-person 2d pose estimation using part affinity fields," in *IEEE/CVF CVPR*, 2017.

[2] S. e. a. Kreiss, "Pifpaf: Composite fields for human pose estimation," in *IEEE/CVF CVPR*, 2019.

[3] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *IEEE/CVF CVPR*, 2019.

[4] C. e. a. Zheng, "3d human pose estimation with spatial and temporal transformers," in *IEEE/CVF ICCV*, 2021, pp. 11 656–11 665.

[5] A. e. a. Zeng, "Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach," in *ECCV*. Springer, 2020, pp. 507–523.

[6] R. A. e. a. Güler, "Densepose: Dense human pose estimation in the wild," in *IEEE/CVF CVPR*, 2018, pp. 7297–7306.

[7] K. e. a. Lin, "End-to-end human pose and mesh reconstruction with transformers," in *IEEE/CVF CVPR*, 2021, pp. 1954–1963.

[8] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *IEEE/CVF ICCV*, 2019.

[9] Z. Li, J. Liu, Z. Zhang, S. Xu, and Y. Yan, "Cliff: Carrying location information in full frames into human pose and shape estimation," in *ECCV*. Springer, 2022, pp. 590–606.

[10] M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black, "PARE: Part attention regressor for 3D human body estimation," in *IEEE/CVF ICCV*. IEEE, 2021, pp. 11 127–11 137.

[11] D. e. a. Xiang, "Monocular total capture: Posing face, body, and hands in the wild," in *IEEE/CVF CVPR*, 2019, pp. 10 965–10 974.

[12] M. J. Black, P. Patel, J. Tesch, and J. Yang, "Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion," in *IEEE/CVF CVPR*, 2023, pp. 8726–8737.

[13] W.-L. Wei, J.-C. Lin, T.-L. Liu, and H.-Y. M. Liao, "Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video," in *IEEE/CVF CVPR*, 2022.

[14] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *IEEE/CVF CVPR*, 2020.

[15] R. Bashirov, A. Ianina, K. Iskakov, Y. Kononenko, V. Strizhkova, V. Lempitsky, and A. Vakhitov, "Real-time rgbd-based extended body pose estimation," in *IEEE/CVF WACV*, 2021, pp. 2807–2816.

[16] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *IEEE/CVF CVPR*, 2019.

[17] S. Dwivedi, Y. Sun, P. Patel, Y. Feng, and M. Black, "Tokenhmr: Advancing human mesh recovery with a tokenized pose representation," in *IEEE/CVF CVPR*, 2024, pp. 1323–1333.

[18] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *IEEE/CVF CVPR*, 2018.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[20] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *ECCV*, 2018.

[21] M. Kaufmann, J. Song, C. Guo, K. Shen, T. Jiang, C. Tang, J. J. Zárate, and O. Hilliges, "EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild," in *IEEE/CVF ICCV*, 2023.

[22] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *ECCV*. Springer, 2016, pp. 561–578.

[23] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *SIGGRAPH Asia*, vol. 34, no. 6, pp. 248:1–248:16, 2015.

[24] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, "Neural body fitting: Unifying deep learning and model based human pose and shape estimation," in *3DV*, 2018, pp. 484–494.

[25] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3d and 2d human representations," in *IEEE/CVF CVPR*, 2017. [Online]. Available: http://up.is.tuebingen.mpg.de

[26] S. Dwivedi, N. Athanasiou, M. Kocabas, and M. Black, "Learning to regress bodies from images using differentiable semantic rendering," in *ICCV*, 2021, pp. 11 230–11 239.

[27] Y. Sun, Q. Bao, W. Liu, Y. Fu, M. J. Black, and T. Mei, "Monocular, one-stage, regression of multiple 3d people," in *IEEE/CVF ICCV*, 2021, pp. 11 159–11 168.

[28] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M. J. Black, "Putting people in their place: Monocular regression of 3d people in depth," *IEEE/CVF CVPR*, pp. 13 233–13 242, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245144814

[29] H. Choi, G. Moon, J. Park, and K. M. Lee, "Learning to estimate robust 3d human mesh from in-the-wild crowded scenes," in *IEEE/CVF CVPR*, 2022, pp. 1465–1474.

[30] S. Dwivedi, C. Schmid, H. Yi, M. Black, and D. Tzionas, "Poco: 3d pose and shape estimation with confidence," in *3DV*, 2024, pp. 85–95.

[31] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *IEEE/CVF CVPR*, 2019, pp. 4496–4505.

[32] K. Lin, L. Wang, and Z. Liu, "Mesh graphormer," in *IEEE/CVF ICCV*. IEEE Computer Society, 2021, pp. 12 919–12 928. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01270

[33] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, "Metrabs: Metric-scale truncation-robust heatmaps for absolute 3d human pose estimation," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, pp. 16–30, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:220514499

[34] M. Mihajlovic, S. Saito, A. Bansal, M. Zollhoefer, and S. Tang, "Coap: Compositional articulated occupancy of people," in *IEEE/CVF CVPR*. IEEE Computer Society, 2022, pp. 13 191–13 200. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01285

[35] S. Saito, T. Simon, J. Saragih, and H. Joo, "Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization," in *IEEE/CVF CVPR*. IEEE Computer Society, 2020, pp. 81–90. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00016

[36] Y. Xiu, J. Yang, X. Cao, D. Tzionas, and M. J. Black, "Econ: Explicit clothed humans optimized via normal integration," in *IEEE/CVF CVPR*. IEEE Computer Society, 2023, pp. 512–523. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00057

[37] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, "Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation," *IEEE/CVF CVPR*, pp. 3382–3392, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:227228054

[38] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3d human pose reconstruction," in *IEEE/CVF CVPR*, 2015, pp. 1446–1455.

[39] G. Georgakis, R. Li, S. Karanam, T. Chen, J. Kosecka, and Z. Wu, "Hierarchical kinematic human mesh recovery," *ArXiv*, vol. abs/2003.04232, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:212633946

[40] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," *IEEE/CVF CVPR*, pp. 10 967–10 977, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:109932872

[41] N. Kolotouros, G. Pavlakos, D. Jayaraman, and K. Daniilidis, "Probabilistic modeling for human mesh recovery," *IEEE/CVF ICCV*, pp. 11 585–11 594, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237303794

[42] R. A. Guler and I. Kokkinos, "Holopose: Holistic 3d human reconstruction in-the-wild," in *IEEE/CVF CVPR*, 2019, pp. 10 884–10 894.

[43] H. Zhu, X. Zuo, S. Wang, X. Cao, and R. Yang, "Detailed human shape estimation from a single image by hierarchical mesh deformation," in *IEEE/CVF CVPR*, 2019, pp. 4491–4500.

[44] H. Zhu, X. Zuo, H. Yang, S. Wang, X. Cao, and R. Yang, "Detailed avatar recovery from single image," *IEEE Trans. on PAMI*, vol. 44, no. 11, pp. 7363–7379, 2022.

[45] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *IEEE/CVF CVPR*, 2017, pp. 109–117.

[46] H. Zhou, D. Greenwood, and S. Taylor, "Self-supervised monocular depth estimation with internal feature fusion," in *BMVC*, 2021.

[47] P. Patel, C.-H. P. Huang, J. Tesch, D. T. Hoffmann, S. Tripathi, and M. J. Black, "AGORA: Avatars in geography optimized for regression analysis," in *IEEE/CVF CVPR*, 2021.

[48] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik, "Humans in 4d: Reconstructing and tracking humans with transformers," in *IEEE/CVF ICCV*, 2023, pp. 14 783–14 794.

[49] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *IEEE/CVF CVPR*, 2020.

[50] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *IEEE/CVF CVPR*, 2019.

[51] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *TPAMI*, 2019.

[52] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[53] Y. Wang and K. Daniilidis, "Refit: Recurrent fitting network for 3d human recovery," in *IEEE/CVF ICCV*, 2023.

[54] C.-H. P. Huang, H. Yi, M. Höschle, M. Safroshkin, T. Alexiadis, S. Polikovsky, D. Scharstein, and M. J. Black, "Capturing and inferring dense full-body human-scene contact," in *IEEE/CVF CVPR*, 2022, pp. 13 274–13 285.

[55] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020.

[56] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[57] H. Kato, Y. Ushiku, and T. Harada, "Neural 3d mesh renderer," in *IEEE/CVF CVPR*, 2018.