DeforHMR: Vision Transformer with Deformable Cross-Attention for 3D Human Mesh Recovery

Jaewoo Heo*1
Stanford University
jeffheo@stanford.edu

George Hu*1
Stanford University
gehu@stanford.edu

Zeyu Wang
Stanford University
wangzeyu@stanford.edu

Serena Yeung-Levy
Stanford University
syyeung@stanford.edu

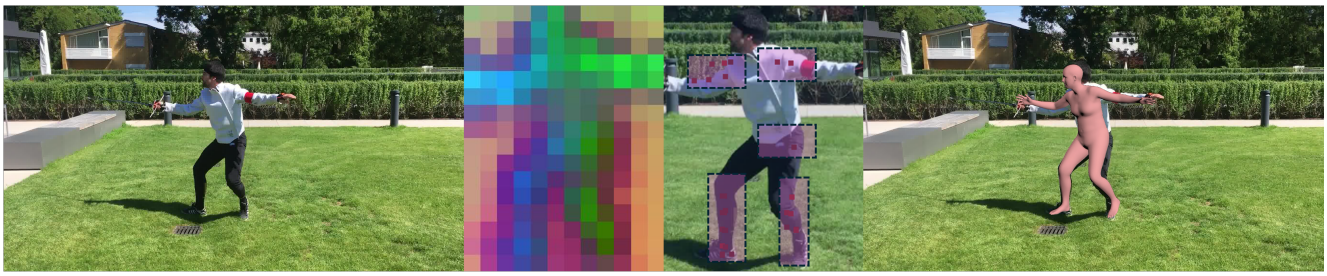

Figure 1. We present **DeforHMR**, a single-image, regression-based methodology for HMR. **DeforHMR** uses a vision transformer (ViT) encoder to derive spatial features from the input image and a deformable cross-attention transformer decoder to learn meaningful spatial relations from the features, enabling the ability to recover accurate 3D human body meshes. **Left to right: (1)** An image in the "outdoors_fencing_01" class of 3DPW [44]. **(2)** Spatial feature of the image outputted by our ViT encoder. **(3)** Attention visualization for the first head of the first layer of our deformable attention transformer decoder. Red and maroon square dots are attention locations; we emphasize the boxed and highlighted areas. **(4) DeforHMR**'s output mesh projected onto original image.

Abstract

*Human Mesh Recovery (HMR) is an important yet challenging problem with applications across various domains including motion capture, augmented reality, and biomechanics. Accurately predicting human pose parameters from a single image remains a challenging 3D computer vision task. In this work, we introduce **DeforHMR**, a novel regression-based monocular HMR framework designed to enhance the prediction of human pose parameters using deformable attention transformers. **DeforHMR** leverages a novel query-agnostic deformable cross-attention mechanism within the transformer decoder to effectively regress the visual features extracted from a frozen pretrained vision transformer (ViT) encoder. The proposed deformable cross-attention mechanism allows the model to attend to relevant spatial features more flexibly and in a data-dependent manner. Equipped with a transformer decoder capable of spatially-nuanced attention, **DeforHMR** achieves state-of-the-art performance for single-frame regression-based methods on the widely used 3D HMR benchmarks 3DPW and RICH. By pushing the boundary on the field of 3D human mesh recovery through deformable attention, we introduce an new, effective paradigm for decoding local spatial information from large pretrained vision encoders in computer vision.*

1. Introduction

Motion capture (MoCap) technology has applications in numerous fields such as film, gaming, AR/VR, as well as sports medicine by providing a tool to capture and analyze human pose in 3D. Traditional marker-based MoCap systems utilizing multi-view cameras and marker suits recover highly accurate human pose but suffer from poor accessibility due to the high cost of setting up the adequate labora-

tory environment [31]. In contrast, a single camera with the correct algorithm can perform 3D Monocular Human Mesh Recovery (HMR), which recovers a mesh of a human body in 3D given an input image or video as a more accessible alternative using deep neural networks [19].

A common parametric approach to 3D HMR leverages the Skinned Multi-Person Linear (SMPL) [29] representation model that regresses joint articulations (often referred to as the pose parameter) and a body shape parameter to generate accurate 3D human body meshes. Current challenges in HMR include occlusion situations [21] and the complexity and variability of human pose, but underlying these issues is simply insufficient spatial understanding in neural networks to output correct pose parameters [26].

More recently, advances in vision transformers [11] have demonstrated versatility and overall impressive performance across a wide range of vision tasks and domains [20], particularly in determining complex spatial relations [14]. In the field of object detection, deformable attention [54] [46] has emerged as one promising solution for accurate, space-aware localization, and extending such an approach to HMR requires even greater focus on extracting precise positional semantics [54].

In parallel, issues of data generalization across diverse real world applications for vision models have been diminished by the release of large vision transformer models pretrained on self-supervision tasks on web-scale datasets [33] [3]. The ability of these foundation models to generate meaningful features across all data spectra for downstream application has created a new effective learning paradigm, and more recently, works [12] have begun on improving the spatial resolution of these vision foundation model features for ever better results. For HMR, [14] has noted how pretraining with both masked auto-encoding [15] along with 2D keypoint prediction [49] [4] has been essential to model convergence, and we build upon along this line of work, investigating how to most effectively decode the features from these large-scale pretrained models.

Integrating the information derived from large, pretrained vision transformer [49] features and deformable attention decoding, we present **DeforHMR**, a novel transformer-based HMR framework that significantly improves upon current methods in both accuracy and computation efficiency.

We believe **DeforHMR** offers significant benefit through the synergy between the semantically-meaningful spatial features from a pretrained vision transformer and the deformable attention mechanism; in deformable attention the reference and offset locations are floating point values in the feature map coordinate space, and bilinear interpolation is used to extract the relevant key and value information. Hence, the advantage of such deformable attention mechanism is derived mostly from data-dependent spatial flexibility, or the ability to dynamically shift attention to relevant spatial regions based on the characteristics of the input feature. We believe utilizing rich features from the transformer encoder would enable the spatial dynamism of the deformable attention mechanism to be influential by learning better spatial relations.

In summary, our contributions are twofold:

1. We present **DeforHMR**, a regression-based monocular HMR framework that demonstrates SOTA performance on multiple well-known public 3D HMR datasets under the single-frame, regression-based setting.
2. Inspired by [46], we propose a novel deformable cross-attention mechanism designed to be query-agnostic and spatially flexible.

2. Related Work

2.1. Monocular HMR

Early work in 3D HMR [34][2] revolved mainly around fitting the SMPL parametric body model to minimize the discrepancy between its reprojected joint locations and 2D keypoint predictions on the 2D image. End-to-end 3D human mesh recovery, not relying upon intermediate 2D keypoints or joints, was first proposed by Kanazawa et al. [19]. This was achieved by leveraging novel deep learning advancements at the time and regressing the SMPL parameters along with a camera model to derive the 3D human meshes. Ever since then, various neural network-based HMR methodologies have been proposed. In [25], the authors aim to resolve the discrepancy between plausible human meshes and accurate 3D key-point predictions through a hybrid inverse kinematics solution involving twist-and-swing decomposition. Li et al. [26] proposes to mitigate the loss of global positional information after cropping the human body through utilizing more holistic features containing global location-aware information to ultimately regress the global orientation better, and, Goel et al. [14] would contribute to this through implementing a vision transformer architecture using a single query token fed into the decoder for SMPL parameter predictions. HMR-2.0 established a new competitive baseline on single human mesh recovery, and in particular, they show how their transformer network can encode and decode complex human pose due to their ability to better capture difficult spatial relations.

2.1.1 Optimization-Based HMR

In [25], Li et al. notes how an alternative approach to direct regression of the SMPL parameters is optimization-based HMR [8][48][47][55], which estimates the body pose and shape through an iterative fitting process. For instance, PLIKS [39] fits a linearized formulation of the SMPL model on region-based 2D pixel-alignment, and ReFit [45] itera-

tively projects 2D keypoints in order to effectively generate accurate meshes. However, optimization-based HMR at inference-time does not have any runtime guarantees and often struggles from large runtime due to the iterative refinement process, and thus can be difficult to integrate into real-time application settings.

2.1.2 Temporal HMR

As computational capacity has increased over recent years, the ability to use complete sequences of video frames for human mesh recovery has recently found success. Within this area of temporal HMR [10][30][7][24], Kocabas et al. [22] have proposed adversarial training with a temporal network architecture to learn to discriminate between real and fake pose sequences. Moreover, [52] proposes to mitigate the effects of occlusions in video sequences through infilling occluded humans with a deep generative motion infiller, and [50] utilizes a temporal motion prior [36] to effectively decouple the human motion and camera motion given a video sequence. More recently, Shin et al. [40] have incorporated motion encoding and SLAM [42] approximation, along with model scale in order to achieve obtain state of the art performance for multi-frame inputs.

While this line of work is promising for integrating complete video information in human mesh recovery, we restrict our focus to single-frame inputs so that our method generalizes to individual images.

2.2. Deformable Attention

The Deformable Transformer [54] architecture, first proposed in end-to-end object detection [5], has demonstrated comparable performance to other SOTA methods without needing any hand-designed components commonly used in object detection [35][13]. The deformable attention module is designed for efficiency and complex relational parameterization, having the keys and values be sparsely sampled learned offsets from a reference location determined by a given query. Zhu et al. [54] show that this increases model training and inference speed while also incorporating inductive biases for precise spatial modeling beneficial for object detection.

Yoshiyasu [51] extends this notion of deformable attention to optimization-based non-parametric 3D HMR with the DeFormer architecture, using the joint and shape query tokens at each layer to generate reference points and offsets on multi-scale maps to be used in the attention computation. DeFormer works directly with positional information without the SMPL parameterization for dense mesh reconstruction, and it improves upon previous baselines of similar model size.

In [46], Xia et al., show how previous works for deformable attention, in fact, function more like a deformable convolution [9] without attention interactions between all queries and all keys. They then propose the Deformable Attention Transformer (DAT), a vision transformer backbone using true deformable self-attention. DAT demonstrates its advantage of deformable self-attention for localization-based tasks such as object detection and instance segmentation, outperforming shifted window full self-attention methods [28] on COCO [27]. In their analysis, they suggest that deformable attention consistently attends to more important and relevant areas of the image and feature map compared to full self-attention, confirming that the true deformable attention interactions between queries and keys result in realized performance and interpretable improvements.

3. Methodology

In this section, we delve into the methodologies of each component of **DeforHMR**. More specifically, we discuss using a frozen ViT pretrained on pose estimation as a feature encoder and our novel deformable cross-attention mechanism. Lastly, we touch upon model training specifications.

3.1. Generating Feature Maps

We use the ViT-Pose from Xu et al. [49] as our initial feature encoder. This is a ViT-H with patch size 16 and input size 256 by 192 that is pretrained with masked autoencoding on ImageNet and 2D pose estimation on COCO [49]. We freeze all the weights during training and pass the input image through to generate the features maps. That is, given an input image $x \in \mathbb{R}^{H' \times W' \times 3}$ and a patch size of 16, we represent the spatial output tokens of the encoder f as $f(x) \in \mathbb{R}^{H \times W \times C}$ for $H = H'/16$ and $W = W'/16$. We freeze the weights of the ViT encoder to isolate the contributions of our novel deformable cross-attention decoder, ensuring that any observed performance improvements stem solely from the decoder's ability to refine feature representations rather than changes in the backbone's learned embeddings.

3.2. Deformable Attention Decoder

3.2.1 SMPL Multi-query Transformer Decoder

Our approach can be thought of as using query tokens for SMPL parameters. These are learnable tokens $t \in \mathbb{R}^{25 \times 1024}$, representing 24 pose tokens and 1 shape token, which further incorporate information from the image features through the decoder blocks.

Following the standard paradigm of the transformer decoder [43], each layer of our deformable cross-attention decoder is compromised of a self-attention, a deformable cross-attention, and a feed-forward network. We further elaborate on our deformable cross-attention mechanism in the subsequent section.

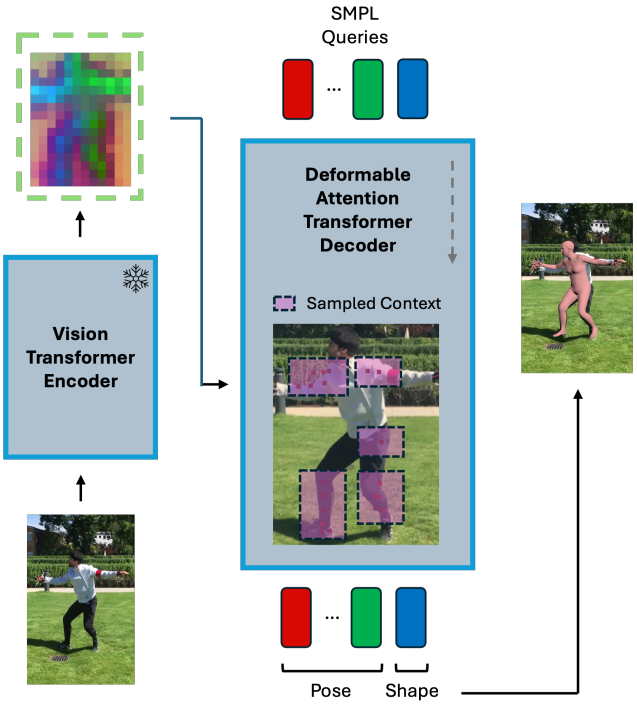

Figure 2. Full system architecture of **DeforHMR**. We dedicate a learnable query embedding for each of the 24 joint articulations and the body shape vector.

After passing the queries through the decoder, we learn linear projections W_{pose} and W_{shape} to get the desired outputs of pose parameters $\theta \in \mathbb{R}^{24\times6}$, and shape parameters $\beta \in \mathbb{R}^{10}$. For the pose rotation angles in the SMPL parameters, we use the common 6D representation proposed by Zhou et al. (2020) [53] for a more continuous loss-landscape, converting to the actual pitch/roll/yaw and rotation matrix afterwards. Moreover, we use one round of iterative error feedback [6], starting with the mean SMPL values from Humans 3.6M [18] to condition our predictions better. These are thus finally passed into the SMPL model to generate our 3D meshes. The 3D meshes are reprojected onto the original image using ground-truth camera parameters provided in each respective dataset.

3.2.2 Deformable Cross-Attention

We propose a novel, query-agnostic deformable cross-attention mechanism designed to capture fine-grained spatial details as shown in 3. The deformable aspect introduces learnable offsets that allow the attention to adaptively select key-value pairs from the feature map, as opposed to uniform attention to all spatial locations. Our method is inspired by the self-attention mechanism proposed in [46].

For some layer ℓ, let the input tokens to this layer be $\mathbf{Y}^{(\ell-1)} \in \mathbb{R}^{B\times25\times D}$, for batch size B, the 25 SMPL tokens, and model dimension $D = 1024$. The first part of our decoder block is multi-head self-attention with residual on the query tokens, so let the output from the self-attention be $\mathbf{Y}_1^{(\ell)} = \text{Self-Attn}(\mathbf{Y}^{(\ell-1)}) + \mathbf{Y}^{(\ell-1)}$ Let the spatial features from the encoder be $\mathbf{X} \in \mathbb{R}^{B\times H\times W\times C}$. We will refer to these spatial features as the *context* for our decoder.

We consider a base set of reference points $\mathbf{R} \in \mathbb{R}^{(BG)\times2\times H\times W}$, which are initialized as grid coordinates normalized to the range $[-1, 1]$. These reference points provide an initial uniform bias for the positions of the keys in the feature map:

$$\mathbf{R}_{ij} = \left(\frac{i}{H} \times 2 - 1, \frac{j}{W} \times 2 - 1\right), \tag{1}$$

where i and j denote the indices over height and width, respectively.

We then compute G unique raw offsets for each reference point by processing the context features through a series of grouped convolutions, non-linear activations (GELU) [16], and normalization layers:

$$\Delta\mathbf{P}' = \text{Conv2D}(\text{GELU}(\text{LayerNorm}(\text{Conv2D}(\mathbf{X})))). \tag{2}$$

These raw offsets are then passed through hyperbolic tangent and scaled by $\frac{\lambda_o}{H_r}$, restricting the offset to within λ_o multiplied by the grid spacing. Hence, the final offsets $\Delta\mathbf{P}$ are

$$\Delta\mathbf{P} = \frac{\lambda_o}{H_r} \tanh(\Delta\mathbf{P}')$$

The resulting offsets $\Delta\mathbf{P} \in \mathbb{R}^{(BG)\times2\times H\times W}$ indicate the amount by which each reference point in the grid of dimension $H \times W$ should be shifted, allowing the model to focus on different parts of the feature map depending on the input context. This mechanism enables the attention to be more flexible and context-aware.

The final sampling positions are hence the sum of the reference point positions and offsets.

$$\mathbf{P} = \Delta\mathbf{P} + \mathbf{R}. \tag{3}$$

With these positions, we sample the context X, employing bilinear interpolation to extract precise embedding values at these adjusted positions. The sampled context $\mathbf{S} \in \mathbb{R}^{(BG)\times\frac{C}{G}\times H\times W}$ can finally be projected into keys $\mathbf{K} = W_k \cdot \mathbf{S}$ and values $\mathbf{V} = W_v \cdot \mathbf{S}$. And likewise, we project the input tokens into the queries $\mathbf{Q} = W_q \cdot \mathbf{Y}_1^{(\ell)}$.

We calculate cross-attention scores by first taking the dot product of the queries $\mathbf{Q}$ with the keys $\mathbf{K}$, then summing this term with an attention bias term, which is computed through sampling a learnable relative positional embedding tensor $\mathbf{Z}_{rpe}$ via bilinear interpolation $\phi(\_, \_)$ using $\mathbf{P}$ as sampling location:

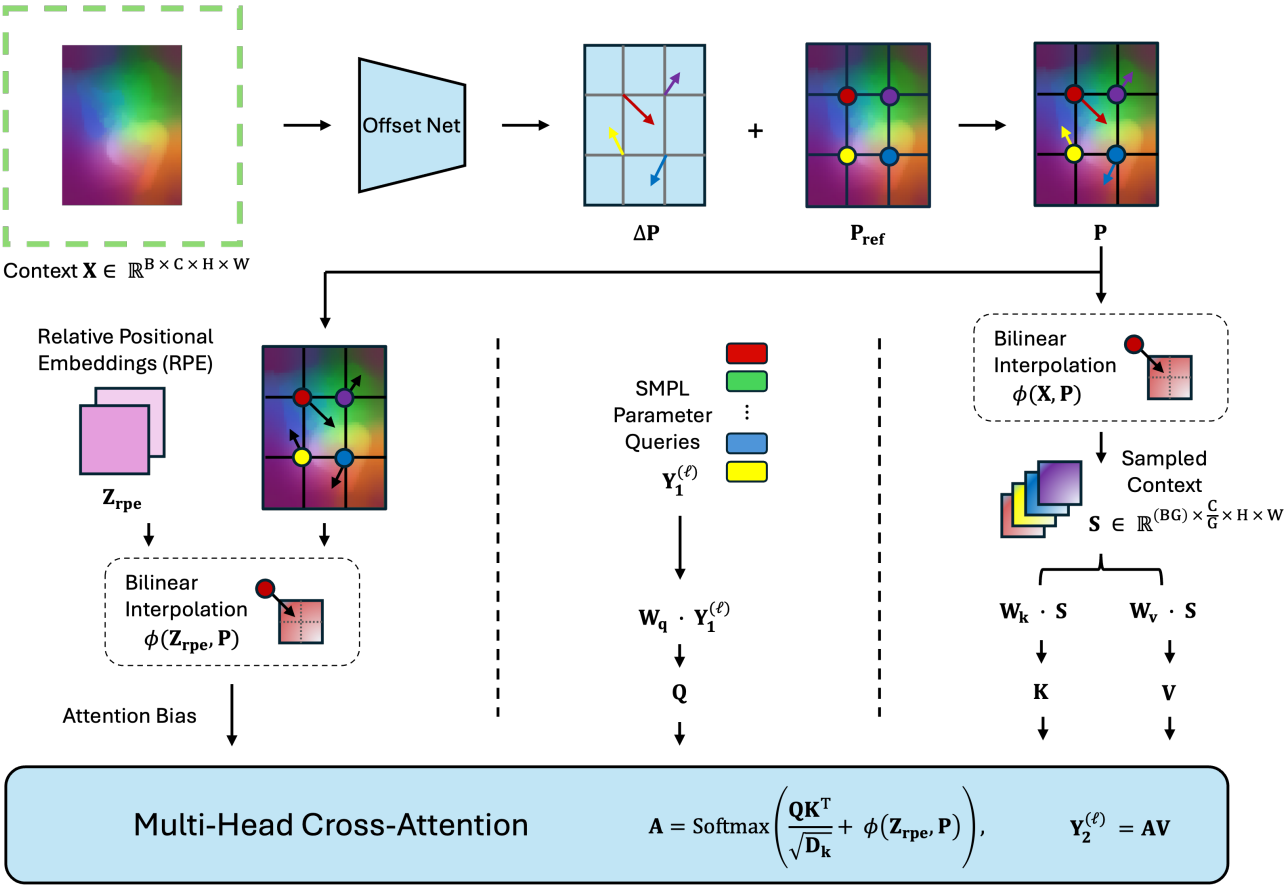

Figure 3. Our proposed deformable cross-attention module. The offset-generating convolutional neural network takes the spatial features from the transformer encoder to generate sampling position offsets ΔP. These offsets are added to the grid reference positions P_{ref} to get our final sampling positions P. These sampling positions are used to 1) sample the input context via bilinear interpolation, which is then projected to keys and values for attention computation, as well as 2) sample the relative positional embeddings (RPE) to get our attention bias. These are combined in the standard multi-head cross-attention formulation with relative position biases to generate the output.

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{D_k}} + \phi(\mathbf{Z}_{rpe}, \mathbf{P})\right) \tag{4}$$

where D_k is the dimension of the keys, used to scale the scores and stabilize training.

More specifically, the relative positional embedding is a unique learned grid for each query position. This is similar to both DAT [46] and the original relative position encoding [37]; the keys have actual positions P, so it uses the grid formulation as in DAT, but the queries do not have positions, so we simply index and learn the relative position embedding separately for each query.

We lastly multiply the attention coefficients by the values and add the residual to get the cross attention output $\mathbf{Y}_2^{(\ell)} = \mathbf{AV} + \mathbf{Y}_1^{(\ell)}$. Passing this through a 2-layer feed-forward network, we get the layer's final output $\mathbf{Y}^{(\ell)} = \text{FFN}(\mathbf{Y}_2^{(\ell)}) + \mathbf{Y}_2^{(\ell)}$

3.2.3 Difference Between DeforHMR and Previous Deformable Cross-Attention

We want to emphasize that our deformable cross-attention differs greatly from DeformableDETR[54]-style cross-attention proposed by previous works. Unlike typical deformable attention methods where offsets are conditioned on the queries, our model computes offsets directly from the context, meaning they are query-agnostic. These query-agnostic offsets are then used to sample the context from our spatial features that would be shared by all queries. This design choice is inspired by the Deformable Attention Transformer [46] (DAT) paper; however, their focus on encoder architectures means their deformable self-attention models do not fully decouple relations between queries and key-values. By having query-agnostic cross-attention here, we can ensure that the shifts in receptive fields and sampling clusters via deformable attention are consistent and coordinated across all queries, capturing global information more

effectively.

3.3. Training Details

Following [26], we train with reconstruction loss on the SMPL parameters, the 3D joint positions, the 3D mesh vertices, and the projected 2D joint positions, all using mean square error. The relative loss weight for SMPL parameters is $\lambda_{SMPL} = 1$, 2D and 3D joint positions is $\lambda_{joint} = 5$, and mesh vertices $\lambda_{mesh} = 60$. For all training runs, we freeze the ViT-Pose to explore efficient decoding methods.

We train all models on real world datasets, two with 3D SMPL ground truth derived from motion capture—3DPW [44] and MPI-INF-3DHP [32]—and three psuedo-labeled from 2D pose ground truth using the CLIFF-annotator [26]: COCO [27], MPII [1], and Humans3.6M [18]. We train for 100 epochs. The evaluation is performed on the test split of 3DPW and RICH, and we use mean per-joint position error (MPJPE), procrustes analysis MPJPE (PA-MPJPE), and per-vertex error (PVE), all in millimeters (mm) to determine how well the model recovers accurate human pose in 3D.

4. Results

We compare various model architectures and approaches using our evaluation metrics in Table 1. Note that since we are interested in single-frame inputs and inference in real-time applications, we exclude multi-frame temporal approaches and optimization-based approaches.

Upon comparing HMR evaluation metrics with several state-of-the-art regression-based HMR methodologies, we demonstrate that **DeforHMR** establishes a new state-of-the-art benchmark on both 3DPW [44] and RICH [17] datasets by a considerable margin.

4.1. Analysis

Our HMR model exhibits a robust capability in capturing the general body pose and proportions of individuals across various scenarios, as seen in the visualizations on Figure 4. Upon rendering the recovered meshes on four distinct images from the test set of the 3DPW [44] and RICH[17] dataset, we confirm our model comprised of the ViT-Pose transformer encoder and the transformer decoder using deformable cross-attention generalizes well across various in-the-wild image examples. Our model demonstrates accurate, plausible, and realistic meshes for humans in various scenarios such as but not limited to executing a fencing motion, walking while conversing sideways, sitting at a table, crouching downwards, etc. In particular, compared to pre-existing HMR models, namely HMR2.0 [14], we show strengths in accuracy of upper body articulation and orientation, as well as feet and hand position.

In Table 2, we decouple some of the main differences between HMR2.0† and **DeforHMR**: multi-query decoder and deformable cross-attention. To do so, we evaluate all four combinations of 1) multi-query versus single-query, and 2) deformable cross-attention versus regular cross-attention on the test set of 3DPW [44]. The ablation results clearly indicate that both the use of multiple queries and the deformable cross-attention mechanism in **DeforHMR** contribute significantly to performance improvements across all three metrics. Specifically, models incorporating these components

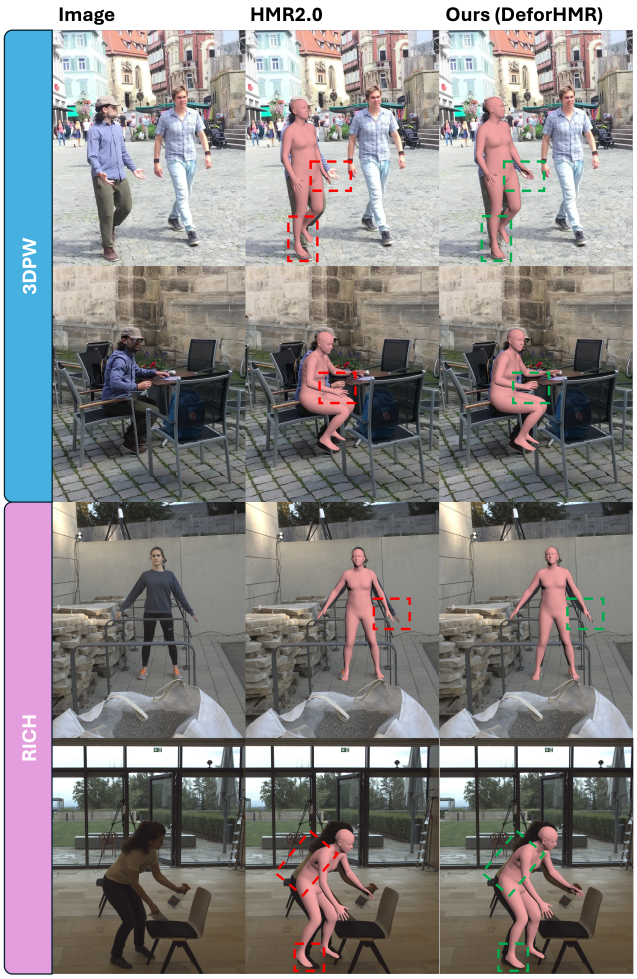

Figure 4. Qualitative results on test set. We visualize the original image and the predicted mesh projected onto the original image for both HMR2.0 [14] and **DeforHMR**. We highlight the inaccurate mesh regions outputted by HMR2.0 in red boxes and highlight the corresponding mesh region on **DeforHMR**'s mesh output in green boxes. Upon visualizing HMR2.0 and our model's recovered meshes on four distinct scenarios across 3DPW [44] and RICH [17], we observe **DeforHMR**'s significant improvements on HMR2.0's ability to recover accurate 3D human meshes. While HMR2.0 shows inaccurate feet & hand positions in all four rows as well as inaccurate orientation of the entire torso in the lowest row, **DeforHMR** consistently shows more accurate feet, hand, and torso positions.

Method	*3DPW* [44]			*RICH* [17]		
	PA-MPJPE ↓	MPJPE ↓	PVE ↓	PA-MPJPE ↓	MPJPE ↓	PVE ↓
ROMP [41]	47.3	76.7	93.4	-	-	-
PARE [23]	46.5	74.5	88.6	60.7	109.2	123.5
CLIFF [26]	43.0	69.0	81.2	56.6	102.6	115.0
HybrIK [25]	41.8	71.6	82.3	56.4	96.8	110.4
SA-HMR [38]	-	-	-	-	93.9	103.0
HMR2.0* [14]	44.4	69.8	83.2	**48.1**	96.0	110.9
DeforHMR	**38.3**	**63.6**	**75.2**	48.6	**84.2**	**94.5**

Table 1. Comparison of state-of-the-art models on 3DPW [44] and RICH [17] datasets. **DeforHMR** achieves superior HMR performance by a wide margin across all metrics on both datasets except PA-MPJPE on RICH that is comparable to that of HMR2.0 [14]. (*) represents the exclusion of 3DPW data during training.

consistently achieve lower error rates, indicating the effectiveness of each design choice, and furthermore, the performance increase going from single to multi-query for deformable cross-attention is much larger than regular cross-attention (4.0mm PVE decrease versus 3.1). This suggests true synergy between the multi-query formulation and deformable cross-attention, enabling superior 3D HMR performance.

Configuration	PA-MPJPE	MPJPE	PVE
Reg-S (HMR2.0†)	41.5	67.2	81.2
Reg-M	39.3	65.3	78.1
Def-S	41.1	66.4	79.5
Def-M (**DeforHMR**)	**38.3**	**63.6**	**75.2**

Table 2. Performance on 3DPW [44] for 1) single-query versus multi-query and 2) regular cross-attention versus our proposed deformable cross-attention. Models with "-S" are single-query models, and models with "-M" are multi-query. "Reg" models use decoders with regular cross-attention, while "Def" employ our deformable cross-attention. Reg-S represents our reimplementation of HMR2.0 with our training data and losses (which we call HMR2.0†), and Def-M is **DeforHMR**.

In Figure 5, we visualize what the first attention head of the first (left) and last (right) layer in **DeforHMR** decoder attends towards. The maroon and red square dots are context positions where the attention value sum over the queries is over 0.25, with bright red corresponding to the largest values. **DeforHMR** is able to incorporate specific information of each individual's limb positions well through the deformable cross-attention, and we can see that the most important positions correspond well with the attention values and locations. Specifically, in the second and third row the model is able to attend towards uncommon arm and leg positions accurately.

Figure 5. Deformable Attention visualizations for first (**left**) and last (**right**) layer of decoder. In each pair of images, we can see important body areas where **DeforHMR** focuses on in order to model these challenging poses.

# Attn Heads	# Groups (G)	Offset Range (λ_o)	*3DPW* [44]			*RICH* [17]		
			PA-MPJPE	MPJPE	PVE	PA-MPJPE	MPJPE	PVE
16	8	1	**38.3**	63.6	75.2	48.6	**84.2**	**94.5**
16	8	2	38.5	63.5	75.4	49.0	84.4	95.4
8	4	1	39.2	63.2	75.5	49.0	85.8	96.2
8	4	2	38.6	**62.5**	**74.5**	**48.3**	85.3	95.8

Table 3. Comparison of **DeforHMR** model performance on 3DPW and RICH datasets across various configurations of deformable cross-attention. We vary the 1) number of attention heads, 2) number of offset groups, and 3) offset range factor. We keep the ratio between the number of attention heads and number offset groups the same in order to scale the information capacity of each offset and head equally.

5. Conclusion

Through this work, we push the boundaries of HMR by combining a pretrained vision transformer encoder with novel deformable cross attention. We have two main contributions. First, we introduce **DeforHMR**, a regression-based framework for monocular HMR that demonstrates SOTA performance on popular 3D HMR datasets such as 3DPW [44] and RICH [17]. Second, inspired by the self-attention mechanism proposed in the Deformable Attention Transformer (DAT) [46], we extend their method with an innovative deformable cross-attention transformer decoder. This mechanism is both query-agnostic and spatially adaptive, enabling the model to dynamically shift focus on relevant spatial features. We show our decoder performs well without additional encoder fine-tuning, allowing for this method to be applicable for API-based large scale models as well.

Despite these advancements, our work possesses some limitations. While our approach demonstrates significant improvements, there is always room for enhancing the model's robustness, particularly towards examples that are inherently more challenging due occlusions and varying lighting conditions in real-world in-the-wild scenarios. Notably, occlusion from obstacles as well as self-occlusion presents a challenge for the model, particularly noticeable in scenarios where one limb occludes another, such as arms or legs during walking motions. These situations often result in inaccurate limb positioning.

Our work reveals several promising future directions, both within HMR and in other applications. Given the effectiveness of deformable cross-attention for decoding information from spatial features, we believe this method can easily be applied to lower-level tasks such as object detection, instance segmentation, keypoint detection, and pose estimation. Moreover, a potential avenue for advancement is applying our deformable attention towards video data and temporal HMR, dynamically attending towards relevant temporal frames as needed. All things considered, **DeforHMR** provides a new effective form of decoding spatial features, a paramount necessity in future applications for large pretrained vision models.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 6

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*. Springer International Publishing, 2016. 2

[3] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and

Percy Liang. On the opportunities and risks of foundation models, 2022. 2

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019. 2

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. 3

[6] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback, 2016. 4

[7] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video, 2021. 3

[8] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention, 2020. 2

[9] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks, 2017. 3

[10] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue, 2019. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 2

[12] Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution, 2024. 2

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, 2014. 3

[14] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers, 2023. 2, 6, 7

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 2

[16] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023. 4

[17] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact, 2022. 6, 7, 8, 1

[18] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 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, 2014. 4, 6

[19] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose, 2018. 2

[20] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys*, 54(10s):1–41, 2022. 2

[21] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery, 2022. 2

[22] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation, 2020. 3

[23] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. Pare: Part attention regressor for 3d human body estimation, 2021. 7

[24] Gun-Hee Lee and Seong-Whan Lee. Uncertainty-aware human mesh recovery from video by learning part-based 3d dynamics. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12355–12364, 2021. 3

[25] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation, 2022. 2, 7

[26] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation, 2022. 2, 6, 7

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 3, 6

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 3

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), 2015. 2

[30] Zhengyi Luo, S. Alireza Golestaneh, and Kris M. Kitani. 3d human motion estimation via motion compression and refinement, 2020. 3

[31] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Wentao Zhu, and Yizhou Wang. 3d human mesh estimation from virtual markers, 2024. 2

[32] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 6

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 2

[34] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and

Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image, 2019. 2

[35] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger, 2016. 3

[36] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation, 2021. 3

[37] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations, 2018. 5

[38] Zehong Shen, Zhi Cen, Sida Peng, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Learning human mesh recovery in 3d scenes. *CVPR*, 2023. 7

[39] Karthik Shetty, Annette Birkhold, Srikrishna Jaganathan, Norbert Strobel, Markus Kowarschik, Andreas Maier, and Bernhard Egger. Pliks: A pseudo-linear inverse kinematic solver for 3d human body estimation, 2023. 2

[40] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. Wham: Reconstructing world-grounded humans with accurate 3d motion, 2024. 3

[41] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J. Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people, 2021. 7

[42] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras, 2022. 3

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 3

[44] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 6, 7, 8

[45] Yufu Wang and Kostas Daniilidis. Refit: Recurrent fitting network for 3d human recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14644–14654, 2023. 2

[46] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Dat++: Spatially dynamic vision transformer with deformable attention, 2023. 2, 3, 4, 5, 8

[47] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild, 2018. 2

[48] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum ghuml: Generative 3d human shape and articulated pose models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (Oral)*, pages 6184–6193, 2020. 2

[49] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation, 2022. 2, 3

[50] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild, 2023. 3

[51] Yusuke Yoshiyasu. Deformable mesh transformer for 3d human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17006–17015, 2023. 3

[52] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras, 2022. 3

[53] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks, 2020. 4

[54] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection, 2021. 2, 3, 5

[55] Nikolaos Zioulis and James F. O'Brien. Kbody: Towards general, robust, and aligned monocular whole-body estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2023. 2

DeforHMR: Vision Transformer with Deformable Cross-Attention for 3D Human Mesh Recovery

Supplementary Material

A. Training and Evaluation Details

We train on two NVIDIA TITAN-RTX GPUs with DDP and global batch size 200. We use AdamW optimizer with learning rate 10^{-4} and weight decay 10^{-3}. For both training and evaluation, in each provided scene, we crop the bounding box of each person and resize it to 256 by 192.

B. Additional Ablation Studies

B.1. Effect of positional encoding type

PE Type	*3DPW* [44]	*RICH* [17]
	MPJPE	MPJPE
No PE	65.4	87.0
Absolute PE	64.0	86.8
Relative PE	**63.6**	**84.2**

Table 4. Comparison of **DeforHMR** model performance on 3DPW and RICH datasets for different positional encoding types. We can observe that the relative positional encoding implementation that we implement results in performance gains, particularly for the out of distribution RICH evaluation dataset.

C. More Qualitative Results

We have provided additional qualitative results in the form of human mesh renderings projected onto the original image for several 3DPW [44] and RICH [17] examples. Please refer to "qualitative_results.pdf" to view these results.