

DeformHMR: 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 **DeformHMR** 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.