Supplementary Document for "HumanRAM: Feed-forward Human Reconstruction and Animation Model using Transformers"

ZHIYUAN YU*†, Hong Kong University of Science and Technology, China
ZHE LI*, Huawei, China
HUJUN BAO, State Key Laboratory of CAD&CG, Zhejiang University, China
CAN YANG‡, Hong Kong University of Science and Technology, China
XIAOWEI ZHOU‡, State Key Laboratory of CAD&CG, Zhejiang University, China

In this supplementary document, we provide implementation details, training details, and single-view results.

A IMPLEMENTATION DETAILS

We follow the decoder-only transformer architecture introduced in LVSM [Jin et al. 2024]. We use a patch size of 8×8 for the input and target patchify linear layers. The transformer contains 12 layers with QK-Norm [Henry et al. 2020] for training stability. Each transformer block consists of a multi-head self-attention layer with 16 heads and a two-layered MLP with GeLU [Hendrycks and Gimpel 2016] activation. The hidden dimensions of the attention and MLP layers are 768 and 3072, respectively. Both layers use standard Pre-Layer Normalization [Ba 2016] and residual connections [He et al. 2016]. The Sigmoid activation is applied after the output linear layer. Besides transformer layers, we apply Layer Normalization after patchify linear layers and before unpatchify linear layers as GS-LRM [Zhang et al. 2024]. For SMPL-X [Pavlakos et al. 2019] neural texture, we set the plane resolution as 128 and the dimension as 16. The parameters of neural texture are initialized with the normal distribution of zero mean and 0.1 standard deviation. Similar to LVSM, we adopt FlashAttention-v2 [Dao 2023] in the xFormers [Lefaudeux et al. 2022] library, gradient checkpointing [Chen et al. 2016], and Bfloat16 data type to accelerate training.

For in-the-wild experiments, we first apply PyMAF-X [Zhang et al. 2023] to estimate the SMPL-X from the input image. Since the estimation assumes an orthogonal camera, which differs from the perspective camera used in our setting, we project 3D SMPL-X joints onto the image plane using the orthogonal camera to get RGB aligned 2D joints. Then we apply Perspective-n-Point (PnP) to re-estimate the perspective camera pose using the 2D-3D joint correspondence. Finally, we use RMBG-2.0 [Zheng et al. 2024] to segment and center the foreground human, which is used as the model input.

B TRAINING DETAILS

Efficient Training Strategy. We apply a two-stage training strategy that is widely adopted in recent LRM-based methods [Jin et al. 2024; Wei et al. 2024; Zhang et al. 2024]: pre-training on low-resolution images (256×256) and finetuning on high-resolution images (512×512). Although such a strategy saves computation costs a lot, it is still a huge barrier for researchers with limited resources. To further reduce the computation consumption, we fix the number of input views to 2 in the pre-training phase. We empirically find that the 2-view pre-trained HumanRAM can generalize to different numbers of input views directly. Then we finetune the model to 4-view inputs on a resolution of 512×512.

Other Details. We train HumanRAM on 8 A100 80GB GPUs. We randomly sample 2 input views and 4 in-between target views for pretraining for each subject. The batch size is set to 12 per GPU. We apply a cosine learning rate scheduler with a peak learning rate of 4e-4 and a warm up of 2500 iterations. The 256-size pre-training stage contains 70k iterations and takes about 2 days. We randomly sample 4 input and 4 target views during the finetuning stage. Notably, if the scan has multiple poses, the target-view pose can differ from the input-view pose. The finetuning comprises 10k iterations with a smaller learning rate of 4e-5 and a smaller batch size of 6 per GPU, which takes 4 additional days. We use gradient clipping of 1.0 during training.

For animation experiments, we continuously finetune the reconstruction model on a mixed scan and avatar dataset. For the scan dataset, the sampling process is the same as above. For the avatar dataset, we sample 4 input images at time t and 4 target images randomly selected from $[t-50, t+50]$. The finetuning adopts the same hyperparameters as above and lasts for 4k steps. When training on ZJUMoCap [Peng et al. 2021], we apply ColorJitter augmentation inspired by NNA [Gao et al. 2023] to prevent overfitting.

C RESULTS ON SINGLE-VIEW INPUT

We present the qualitative comparison of single-view animation in Fig. 1 and in-the-wild results in Fig. 2. For in-the-wild experiments, our method can achieve high-quality novel view synthesis for standing humans. However, the animation occurs artifacts when the input human wears an out-of-distribution cloth (e.g., row 3 and 4). We believe our method can address such issues by scaling up training on large-scale real-captured multi-view human datasets like MvHumanNet [Xiong et al. 2024].

*The first two authors contributed equally to this work.
†Work done during an internship at Huawei.
‡Corresponding authors.

Authors' addresses: Zhiyuan Yu, Hong Kong University of Science and Technology, Mathematics, Hong Kong, China, zyuaq@ust.hk; Zhe Li, Huawei, Hangzhou, China, lizhe_thu@126.com; Hujun Bao, State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, China, bao@cad.zju.edu.cn; Can Yang, Hong Kong University of Science and Technology, Mathematics, Hong Kong, China, macyang@ust.hk; Xiaowei Zhou, State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, China, xwzhou@zju.edu.cn.

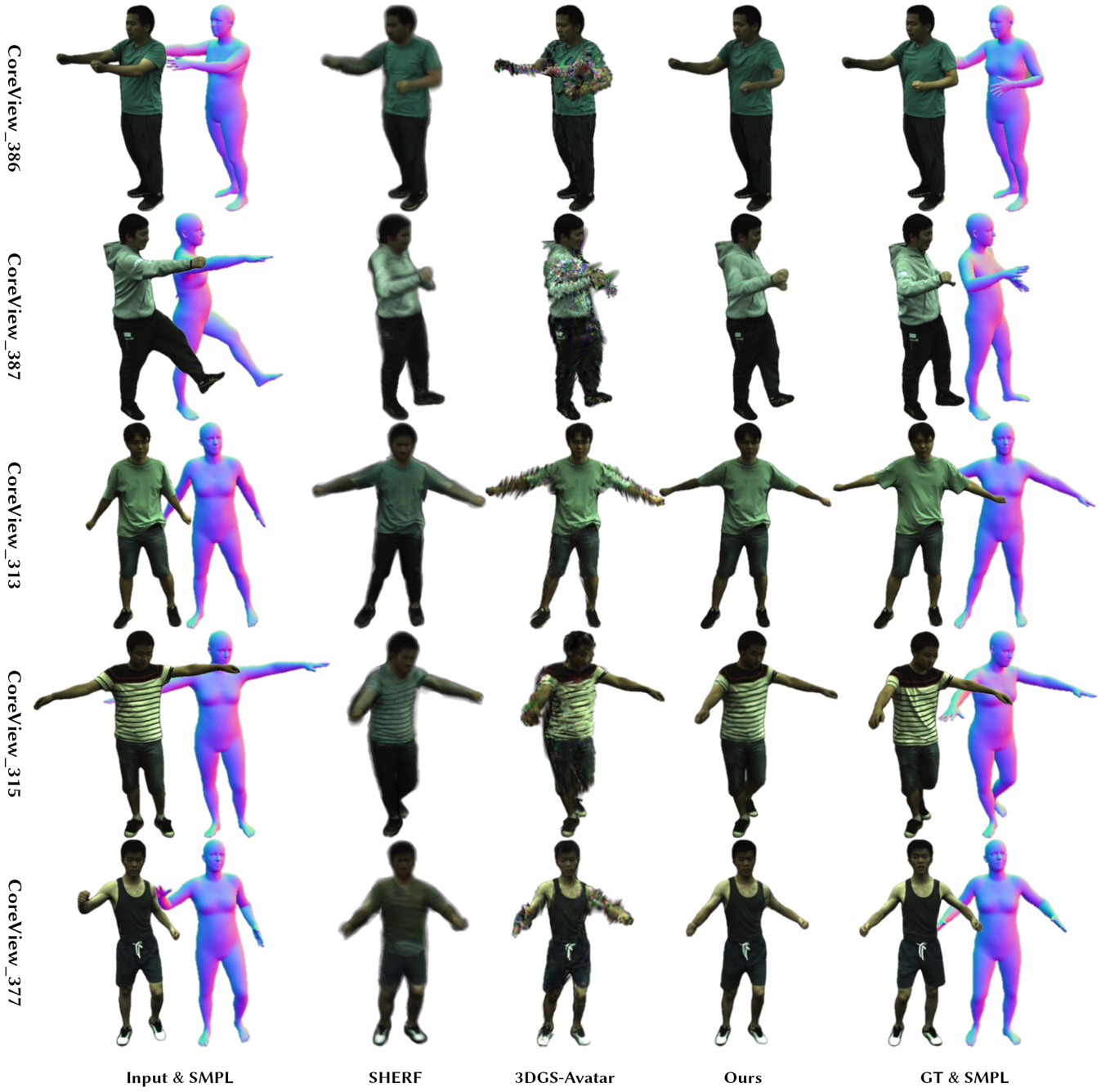

Fig. 1. **Qualitative comparisons for single-view animation on ZJUMoCap [Peng et al. 2021]**. The first two rows are from seen subjects with unseen poses and the last three rows are from unseen subjects.

REFERENCES

Jimmy Lei Ba. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174* (2016).
Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691* (2023).
Qingzhe Gao, Yiming Wang, Libin Liu, Lingjie Liu, Christian Theobalt, and Baoquan Chen. 2023. Neural novel actor: Learning a generalized animatable neural representation for human actors. *IEEE Transactions on Visualization and Computer Graphics* (2023).
Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

Fig. 2. **Qualitative results on in-the-wild images.** We take a monocular image as input and show its reconstruction and animation results. The driving poses are from AMASS [Mahmood et al. 2019].

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).

Alex Henry, Prudhvi Raj Dachapally, Shubham Pawar, and Yuxuan Chen. 2020. Query-key normalization for transformers. *arXiv preprint arXiv:2010.04245* (2020).

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. 2024. Lvsm: A large view synthesis model with minimal 3d inductive bias. *arXiv preprint arXiv:2410.17242* (2024).

Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. 2022. xFormers: A modular and hackable Transformer modelling library. https://github.com/facebookresearch/xformers.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *International Conference on Computer Vision*. 5442–5451.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 10975–10985.

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *CVPR*.

Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. 2024. Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385* (2024).

Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. 2024. MVHumanNet: A Large-scale Dataset of Multi-view Daily Dressing Human Captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19801–19811.

Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. 2023. PyMAF-X: Towards Well-aligned Full-body Model Regression from Monocular Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. 2024. GS-LRM: Large Reconstruction Model for 3D Gaussian Splatting. *European Conference on Computer Vision* (2024).

Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. 2024. Bilateral Reference for High-Resolution Dichotomous Image Segmentation. *CAAI Artificial Intelligence Research* (2024).