

Supplementary Material for: "HMR-Adapter: A Lightweight Adapter with Dual-Path Cross Augmentation for Expressive Human Mesh Recovery"

Anonymous Authors

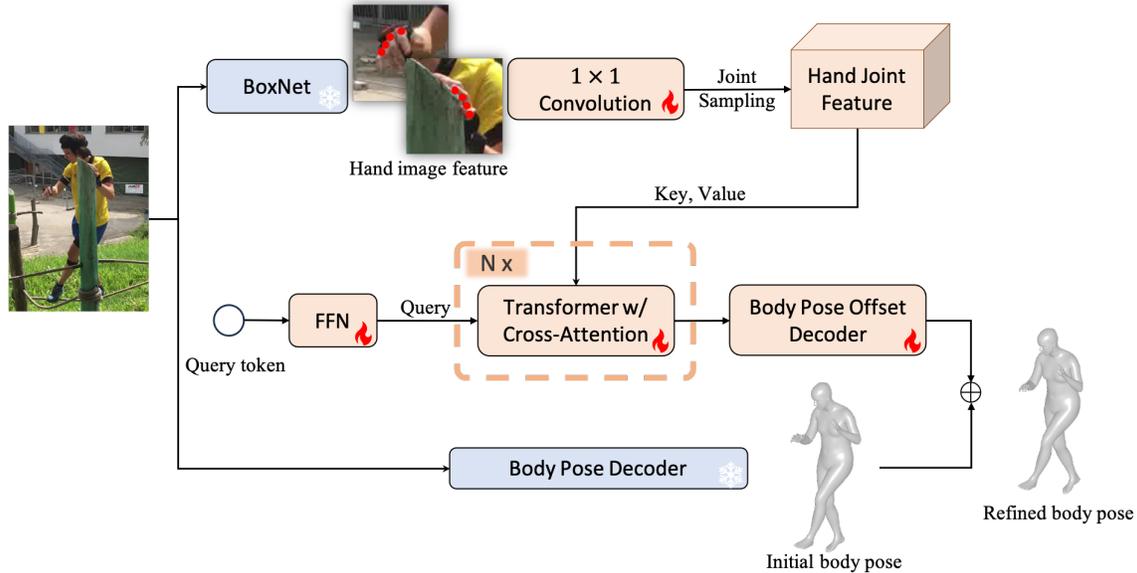


Figure 1: The overview of our HMR-Adapter on the whole-body model. HMR-Adapter refines body pose estimation with additional hand joint features. The frozen whole-body model first estimates the initial body pose and hand bounding boxes. Then the sampled hand joint features are mapped as query key and value into HMR-Adapter’s cross-attention layer. The transformer block is repeated N times before entering a body pose offset decoder. The body pose offset is added to the initial body pose to generate the final refined body pose.

1 IMPLEMENTATION DETAILS

Figs. 1 and 2 illustrate the architectures of HMR-Adapter for the whole-body model and the hand expert model, respectively. In Fig. 1, our HMR-Adapter enhances body pose estimation by integrating additional hand joint features. Initially, a frozen model estimates the initial body pose and hand regions. The hand joint features are obtained by sampling hand image features with the finger root joints. The hand joint features are then processed through HMR-Adapter’s multiple cross-attention layer. Finally, a body pose offset decoder refines the initial estimate, producing the refined body pose. In Fig. 2, HMR-Adapter improves hand pose estimation by adding body pose features. The hand expert model, which remains fixed, includes a transformer-based decoder. We retain this original decoder while adding parallel self-attention and cross-attention layers. Body pose features are processed by the new cross-attention layer. The outputs from both the original and the new cross-attention layers are combined and then sent to the original hand pose decoder to produce a refined hand pose.

During the training of HMR-Adapter for both the hand expert and the whole-body model, we utilize the AdamW optimizer [3] with a learning rate of 1×10^{-5} , $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a

weight decay of 1×10^{-4} . We apply distinct weights to different losses: for HMR-Adapter on the hand expert, the weights for λ_{2D} , λ_{3D} , λ_v , λ_θ , λ_β are set to 0.01, 0.05, 0.001, 0.001, and 0.0005 respectively; for HMR-Adapter on the whole-body model, the weights for λ_{2D} , λ_{3D} , λ_θ , λ_β are 0.01, 0.05, 0.001, and 0.0005 respectively.

2 HMR-ADAPTER VISUALIZATION

To elucidate the functionality of HMR-Adapter on the hand expert model, we visualize its attention maps. In Fig. 3, we show several examples along with their predictions. HMR-Adapter is employed to integrate additional body pose guidance, specifically demonstrating the body pose attention map of HMR-Adapter for the related hand. The visualization reveals that HMR-Adapter selectively emphasizes upper limb poses.

For example, in the instance (d), the elbow and wrist receive higher attention scores, significantly influencing the poses and global orientation of the occluded right hand. This effect can be attributed to the direct impact of the elbow and wrist joints on the global orientation of the hand. When this body pose information is combined with the hand image feature from the frozen hand

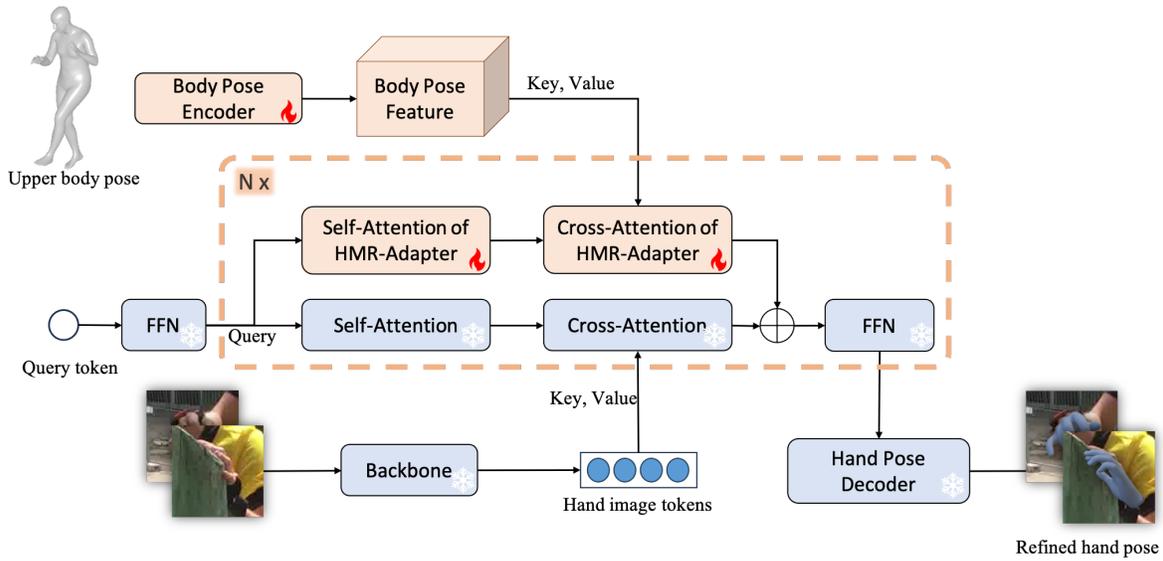


Figure 2: The overview of our HMR-Adapter on the hand expert model. HMR-Adapter injects additional body pose features to refine the hand pose estimation. The frozen hand expert adopts a transformer based decoder. So we keep the original decoder and use a parallel self-attention and cross-attention layer. The new cross-attention layer receives encoded body pose features as input to produce the query key and value matrix. We add the output from the original cross-attention layer to the output from HMR-Adapter’s cross-attention layer. Then the output of the adapted decoder is forwarded to the original hand pose decoder to obtain the refined hand pose.

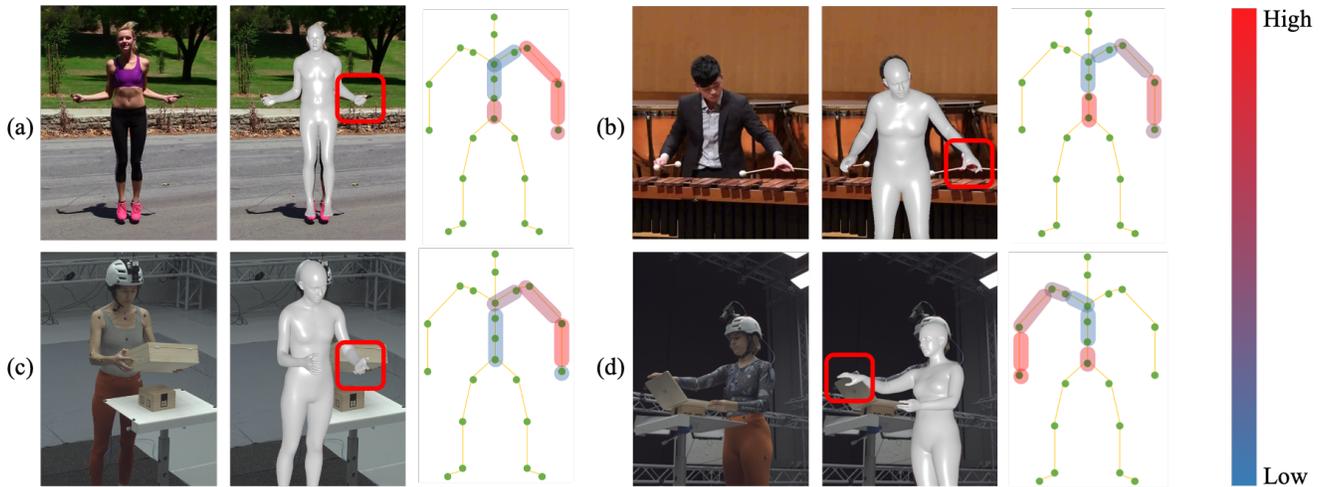


Figure 3: Visualization of HMR-Adapter attention map on the hand expert. It includes the input images, the predicted whole-body meshes with the target hand in a red box, and the corresponding attention maps. (a)~(b) are in-the-wild images from the UBody dataset [2], and (c)~(d) are from the ARCTIC dataset [1].

expert model, the adapted hand model provides robust and plausible estimations for the hand poses.

REFERENCES

[1] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. 2023. ARCTIC: A dataset for dexterous

bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12943–12954.

[2] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. 2023. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21159–21168.

[3] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).