HeatFormer: A Neural Optimizer for Multiview Human Mesh Recovery Supplementary Material

Yuto Matsubara Ko Nishino
Graduate School of Informatics, Kyoto University
https://vision.ist.i.kyoto-u.ac.jp/

1. Implementation Details

HeatEncoder We train HeatEncoder using 1 NVIDIA RTX A6000 with a batch size of 32 and use the AdamW optimizer with a learning rate of 1e-5 for 50 epochs. The learning rate is multiplied by 0.2 each time it reaches 20, 30, and 40 epochs. We use the 3D joint loss, 2D joint loss, and SMPL parameter loss and train on Human3.6M [4] and MPI-INF-3DHP [5] datasets for about 3.5K iterations per epoch. The training data ratio is approximately 2:1. HeatEncoder training takes about three days.

HeatFormer We then freeze HeatEncoder and train the entire HeatFormer using 1 NVIDIA A100 with a batch size of 8 and 4 views for each batch. Same as the HeatEncoder, we use the AdamW optimizer with a learning rate of 1e-5 for 50 epochs and the learning rate is multiplied by 0.2 each time it reaches 30 and 40 epochs. The training dataset ratio is similar to HeatEncoder. HeatFormer training takes about six days.

2. Dataset Details

We describe the details of Human3.6M [4], MPI-INF-3DHP [5], BEHAVE [1] and RICH [3] datasets. The BEHAVE and RICH datasets are used only for testing.

Human3.6M We preprocess the Human3.6M dataset following [2]. Human3.6M does not have ground-truth SMPL parameters. Instead, we use the pseudo ground-truth SMPL parameters generated by NeuralAnnot [6]. We sample the dataset every 20 frames which amounts to about 20K frames of training data for each view.

MPI-INF-3DHP Same as the Human3.6M dataset, we preprocess the data following [2] and use pseudo ground-truth SMPL parameters generated from NeuralAnnot [6]. We removed data whose MPJPE computed on the pseudo ground-truth SMPL parameters exceeds 40mm as they lack reliable ground truth on only train split. We sample every 10 frames which results in about 10K frames for each view.

BEHAVE The BEHAVE dataset is a dataset capturing, with 4 views, human-object interactions in natural environments. We use the BEHAVE dataset to evaluate the geenralization capability and occlusion-robustness of our model. We follow the train and test splits of the BEHAVE dataset and evaluate and compare on the test data. Qualitative results on the BEHAVE dataset are shown in Sec. 3.

RICH The RICH dataset is a real scene dataset taken from 4 views. We show qualitative results on the RICH dataset in Sec. 3.

3. Qualitative Results

We show qualitative results for different datasets, Human3.6M Fig. 5, MPI-INF-3DHP Fig. 6, BEHAVE Fig. 7, and RICH Fig. 8. All results are estimated by HeatFormer trained on the Human3.6M and MPI-INF-3DHP datasets. The results clearly show that HeatFormer is an occlusion-robust, view-flexible, and generaliziable neural optimizer for multiview HMR.

4. Calibration

HeatFormer uses camera extrinsics only for AdaFuse [7] and heatmap projection. HeatFormer can estimate the SMPL parameters without camera extrinsics by skipping AdaFuse and estimating weak-perspective camera parameters instead of the translation calculated with extrinsics. Tab. 9 shows accuracy comparison between HeatFormer applied to calibrated and uncalibrated cameras for 4 views of 3-iter model. As the results show, leveraging calibration information leads to higher accuracy, but even uncalibrated HeatFormer achieves reasonable accuracy.

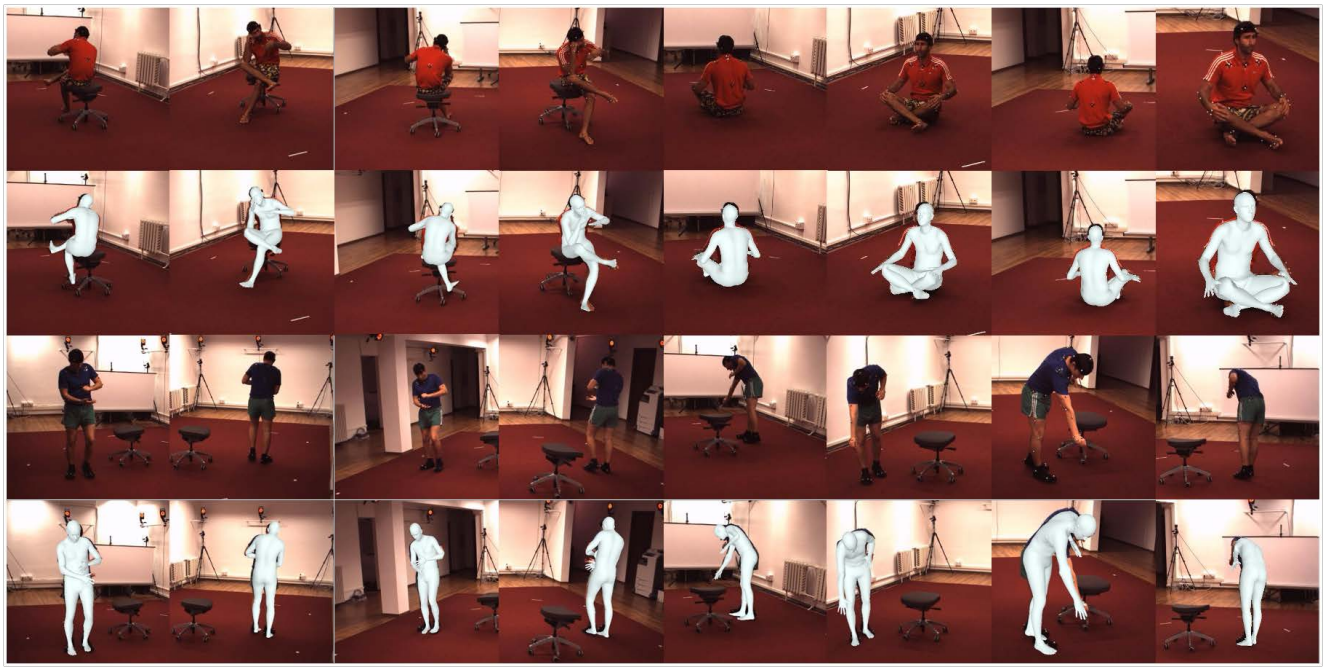

Figure 5. Qualitative results on the Human3.6M [4] dataset. HeatFormer successfully leverages the multiview observations to resolve the complex occlusions.

Figure 6. Qualitative results on the MPI-INF-3DHP [5] dataset. The body shape and pose behind various kinds of occlusions are successfully recovered.

5. Ablation Study

HeatFormer achieves HMR with neural optimization. We use three forward inferences through the decoder (three unrolled iterations) by default. Tab. 10 shows the accuracy for models trained for different numbers of iterations (1 to 5). The results clearly show that the more iterations the better but with diminishing returns. We empirically found three or

Figure 7. Qualitative results on the BEHAVE [1] dataset. This dataset is not used in training. HeatFormer generalizes well to unseen scenes and unseen types of occlusion.

	MPJPE↓	PA-MPJPE↓
calibrated(iter3)	30.3	23.2
uncalibrated(iter3)	42.5	25.8

Table 9. The comparison between calibrated and uncalibrated of 3-iter model.

the # of iterations	MPJPE ↓	PA-MPJPE ↓	MPVPE ↓
1	34.1	27.2	43.9
2	30.7	25.2	39.0
3	28.6	23.1	36.2
4	27.2	22.1	35.2
5	27.5	22.4	34.9

Table 10. Ablation study on the number of forward inferences through the decoder (*i.e.*, number of neural optimization iterations). We train on Human3.6M and MPI-INF-3DHP and test on Human3.6M.

four iterations suffice for the HMR accuracy to converge.

HeatEncoder consolidates the joint heatmaps for each view to extract a rich integrated feature map reflecting the shape and pose of the person observed in the view. This consolidation is essential for the decoder to align the heatmaps through cross-attention and its iterative application to arrive at accurate view-dependent SMPL estimates. The HeatEncoder applied both to the input heatmaps as well as the heatmaps computed from the current SMPL estimate plays a crucial role in making full use of the spatial coordination of joints in each view. We evaluate the effectiveness of HeatEncoder by replacing it with a simple alternative of just taking the max values of the heatmaps across all joints (*i.e.*, per-pixel max-pooling) and applying a heatmap-pretrained ViT to extract features. We train on Human3.6M and MPI-INF-3DHP and test on Human3.6M. The results shown in Tab. 11 clearly demonstrate the effectiveness of HeatEncoder.

	MPJPE ↓	PA-MPJPE ↓	MPVPE ↓
w/HeatEncoder	28.6	23.1	36.2
w/o	30.5	25.6	38.3

Table 11. Evaluation of the effectiveness of HeatEncoder by replacing it with simple max-pooling and ViT feature extraction ("w/o"). The accuracy is evaluated on Human3.6M. The results clearly show that HeatEncoder is essential for heatmap consolidation in each view.

HeatFormer leverages multiview images and updates SMPL parameters with cross-attention. As an abalation study of the decoder of HeatFormer, we replace it with simple pooling and MLP. Tab. 12 shows the results which clearly demonstrate the effectiveness of the architecture of HeatFormer.

Heatmaps computed from the current SMPL model are combined with image features computed from the input

Figure 8. Qualitative results on RICH [3] dataset. This dataset is not used in training. The results clearly demonstrate the strong generalization capability and occlusion-robustness of HeatFormer.

	MPJPE↓	PA-MPJPE↓
w/HeatFormer	28.6	23.1
w/o	60.4	46.4

Table 12. Evaluation of the effectiveness of HeatFormer by replacing it with simple pooling and MLP. The accuracy is calculated on Human3.6M. The results clearly show that the decoder of HeatFormer is essential for neural optimization.

	MPJPE ↓	PA-MPJPE ↓	MPVPE ↓
w/Image Feature	28.6	23.1	36.2
w/o	42.8	33.8	56.1

Table 13. Evaluation of the effectiveness of using image features in combination with heatmaps for the decoder queries. The accuracy is evaluated on Human3.6M. Combining the image features is essential for accurate estimation through decoder inference.

views to form decoder queries for cross-attention with the encoder output tokens. Without these image features, the cross-attention is unlikely to produce meaningful transformations to the heatmaps as the decoder would not know view-correspondences. To confirm this, we evaluate the model without combining image features with the view-dependent heatmaps as decoder queries. Please note that we use image features only for global orientation estimation. We train on Human3.6M and MPI-INF-3DHP, and test on Human3.6M. Tab. 13 show the results which clearly show that the use of image features with the heatmaps is essential for accurate decoder inference.

A key contribution of HeatFormer lies in the adoption of heatmaps as the fundamental representation of pose and their seamless integration in the neural optimization pipeline, which is essential to obtain dense spatial gradi-

	MPJPE↓	PA-MPJPE↓
Heatmap	30.3	23.2
Joint	51.2	37.0

Table 14. Comparison using heatmap representation with direct tokenization of keypoint locations.

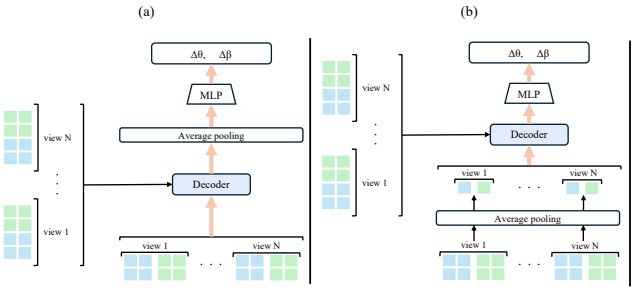

Figure 9. (a) HeatFormer decoder in which the spatial configuration of heatmaps are retained and view-dependent outputs are averaged pooled to compute the SMPL parameters. (b) A variant in which the patchified heatmaps for each view are average pooled before input to the decoder as queries.

	MPJPE ↓	PA-MPJPE ↓	MPVPE ↓
(a)	28.6	23.1	36.2
(b)	36.3	26.6	46.0

Table 15. Comparison of the decoders in Fig. 9. The HeatFormer decoder which retains the spatial configuration of query heatmaps (a) achieves higher accuracy than the variant that consolidates spatial information of the heatmaps through average pooling before they are input to the decoder as queries (b).

ents to let it learn to optimize. To confirm the effectiveness of heatmap representation, we compare with just tokenizing keypoint locations (*i.e.*, directly input and estimate joint coordinates as tokens). Tab. 14 shows that the heatmaps are essential for our high accuracy.

SMPL parameters are computed from the decoder output after average pooling. Retaining the patchified spatial structure of the decoder queries throughout the cross-attention and only average pooling after it is essential to fully leverage the spatial configuration of the heatmaps across the different views (Fig. 9(a)). We confirm the importance of retaining this spatial configuration by comparing it with a variant of the decoder where the patchified heatmaps for each view are average pooled before input to the decoder (Fig. 9(b)) and thus the spatial configuration is dampened. We train on the Human3.6M and MPI-INF-3DHP datasets and test on the Human3.6M dataset. As Tab. 15 shows retaining the spatial configuration of the heatmaps through cross-attention and then consolidating the views via average pooling (*i.e.*, the decoder of HeatFormer) achieves higher accuracy.

	MPJPE ↓	PA-MPJPE ↓	MPVPE ↓
(a)	28.6	23.1	36.2
(b)	36.3	22.8	47.6

Table 16. Comparison between view-dependent estimation (a) and average global estimation (b). Even with the crude averaging, if necessary, HeatFormer's view-dependent estimates can be consolidated without too much loss in accuracy.

HeatFormer outputs view-dependent SMPL estimates, *i.e.*, the SMPL parameters explain each image independently. If a single estimate is necessary, we could combine these view-dependent estimates in any way suitable for the downstream task. For instance, if a SMPL model is necessary for a view in between the input views, SMPL estimates of closest views can be combined. A simple approach to consolidating all views would be to average pool them. We can take the average of the pose parameters without the global orientation and also all the shape parameters. Tab. 16 shows the results of comparing the accuracies of this average global SMPL estimate and our view-dependent estimates. The view-dependent estimates are naturally more accurate, but even a crude averaging will not loose too much accuracy. In general, for downstream tasks, we can select closest views to achieve higher accuracy than averaging.

References

[1] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and Method for Tracking Human Object Interactions. In *CVPR*. IEEE, 2022. 1, 3

[2] 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. In *CVPR*, 2021. 1

[3] 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. In *CVPR*, pages 13274–13285, 2022. 1, 4

[4] 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 TPAMI*, 36(7):1325–1339, 2014. 1, 2

[5] 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 *3DV*. IEEE, 2017. 1, 2

[6] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. NeuralAnnot: Neural Annotator for 3D Human Mesh Training Sets. In *CVPRW*, 2022. 1

[7] Zhe Zhang, Chunyu Wang, Weichao Qiu, Wenhu Qin, and Wenjun Zeng. AdaFuse: Adaptive Multiview Fusion for Accurate Human Pose Estimation in the Wild. *IJCV*, pages 1–16, 2020. 1