

Supplemental Materials: ARTS: Semi-Analytical Regressor using Disentangled Skeletal Representations for Human Mesh Recovery from Videos

Anonymous Authors

This supplemental material contains the following parts:

- (1) The architecture of 3D pose estimation.
 - (2) The derivation of Bone-Guided Shape Fitting.
 - (3) Additional quantitative results.
 - (4) Details about loss functions.
 - (5) Additional visualization results.
- Our code is available at the anonymous page <https://anonymous.4open.science/r/ARTS>.

1 ARCHITECTURE OF 3D POSE ESTIMATION

Figure 1 shows the detailed architecture of the proposed dual-stream 3D pose estimation. Firstly, we utilize the off-the-shelf 2D pose detectors [1, 2] to obtain the 2D skeletons $S^{2D} \in \mathbb{R}^{T \times K \times 2}$ from video frames. Then, we project the 2D skeletons to high-dimensional joint features $X \in \mathbb{R}^{T \times K \times C_1}$. We add the spatial and temporal embeddings to joint features and feed joint features to the dual-stream Transformer network, which consists of a spatial-temporal Transformer and a temporal-spatial Transformer in each block. Subsequently, we add the output of two streams and feed it to the subsequent block. Finally, the joint features after L_1 blocks dual-stream Transformer are regressed from C_1 to 3 to obtain the 3D skeletons $S^{3D} \in \mathbb{R}^{T \times K \times 3}$.

2 DERIVATION OF BONE-GUIDED SHAPE FITTING

SMPL Model. In this work, we employ the SMPL [3] human model for human pose and shape representation. The SMPL model controls the human mesh vertices $M \in \mathbb{R}^{V \times 3}$ through a few number of pose and shape parameters, where $V = 6890$. The pose parameters $\theta \in \mathbb{R}^{72}$ use 3-dimension axis angle to represent relative 3D rotation of $K' = 24$ joints, $\theta = \{\theta_1, \theta_2, \dots, \theta_{K'}\}$. The shape parameters $\beta \in \mathbb{R}^{10}$ are parameterized by the first 10 principal components of the PCA body shape basis. SMPL provides a differentiable blend function $\mathcal{M}(\beta, \theta)$ that maps the pose and shape parameters to the specific human mesh, which can be expressed as follows:

$$\mathcal{M}(\beta, \theta) = \mathcal{F}(\bar{T}_P(\beta, \theta), J(\beta), \theta, \mathcal{W}), \quad (1)$$

$$\bar{T}_P(\beta, \theta) = \bar{T} + B_S(\beta) + B_P(\theta), \quad (2)$$

where $B_S(\beta)$ and $B_P(\theta)$ represent the shape blend and pose blend offsets for the template human mesh \bar{T} , $\mathcal{F}(\cdot)$ is the standard linear blend skinning function, \mathcal{W} is the blend weights pre-trained by large-scale datasets.

Shape Parameters Derivation. In the blend function, the SMPL defines the joint locations in the rest pose as a function of the body shape parameters $J(\beta)$, which is calculated as follows:

$$J(\beta) = W(\bar{T} + B_S(\beta)), \quad (3)$$

$$B_S(\beta) = \sum_{n=1}^{|\beta|} \beta_n S_n, \quad (4)$$

Table 1: Comparison with images-based methods that use 3DPW dataset for training. ‘†’ represents training w/o 3DPW training dataset. The top two best results are highlighted in bold and underlined, respectively.

Method	3DPW			
	MPJPE ↓	PA-MPJPE ↓	PVE ↓	ACCEL ↓
ROMP [4] (ICCV’21)	79.7	49.7	94.7	-
METRO [5] (ECCV’22)	77.1	47.9	88.2	-
CLIFF [6] (ECCV’22)	72.0	45.7	85.3	24.7
MotionBERT [7] (ICCV’23)	76.9	47.2	88.1	-
IKOL [8] (AAAI’23)	73.3	45.5	86.4	-
SimHMR [9] (MM’23)	73.3	<u>45.3</u>	<u>85.4</u>	-
ShapeBoost [10] (AAAI’24)	75.3	44.6	-	-
DPMesh [11] (CVPR’24)	73.6	47.4	90.7	-
ARTS (Ours)†	<u>69.9</u>	48.8	85.6	<u>6.6</u>
ARTS (Ours)	67.7	46.5	81.4	6.5

where $W \in \mathbb{R}^{K \times 6890}$ is the joint regression matrix that obtains joints from human mesh, $\beta = [\beta_1, \dots, \beta_{|\beta|}]^T$, $|\beta| = 10$ is the number of linear shape coefficients. The $S_n \in \mathbb{R}^{3V}$ represents orthonormal principal components of shape displacements. By transforming all the shape displacements to shape blend matrix $S \in \mathbb{R}^{V \times 3 \times 10}$, we can express the joint regression function as:

$$J(\beta) = W(\bar{T} + S\beta). \quad (5)$$

Therefore, to get the bone-aligned shape parameters, we first use the bone length to obtain the bone-aligned joint locations. Then, we use joint locations to derive initial shape parameters, which match the bones and reflect body shape. The equation is as follows:

$$\beta'_{init} = S^{-1}(W^{-1}J_{aligned} - \bar{T}). \quad (6)$$

The Bone-Guided Shape Fitting (BSF) estimates the SMPL shape parameters from the bone length based on Equation 6. However, estimating human shape from sparse bone length is an ill-posed problem. Although the initial SMPL shape parameters contains basic body shape, such as body height and the length of body parts (e.g., legs and arms), it lacks sufficient information about body weight (e.g., fat or thin). Therefore, we further refined the SMPL shape parameters from image features in the Motion-Centric Refinement (MCR) module.

3 ADDITIONAL QUANTITATIVE RESULTS

Comparison with Image-based Methods. Table 1 compares our ARTS with image-based methods on the 3DPW dataset. All methods are trained with the 3DPW dataset. Image-based methods focus on per-frame accuracy and train advanced networks to extract image features with additional 2D datasets, such as HRNet [6]

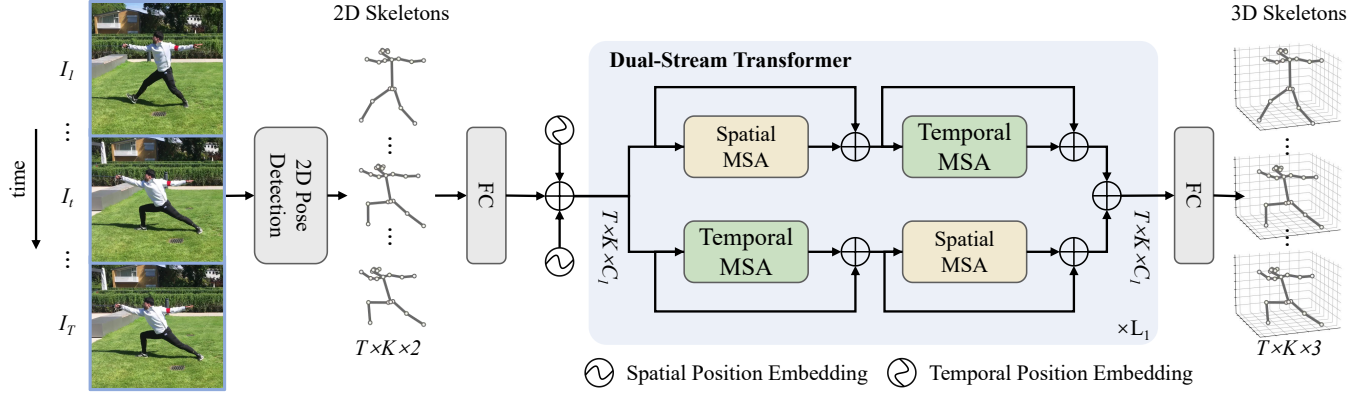


Figure 1: The architecture of 3D pose estimation.

Table 2: Ablation study for human motion padding and bone length on 3DPW.

Method	3DPW			
	MPJPE ↓	PA-MPJPE ↓	MPVPE ↓	Accel ↓
Frame Bone Length	68.6	46.7	82.5	6.7
Zero Motion Padding	68.5	47.0	82.7	7.0
ARTS (Ours)	67.7	46.5	81.4	6.5

and Diffusion network [11], which shows high performance on accuracy. On the contrary, our ARTS utilized the pre-trained CNN backbone [12] to extract image features following previous video-based methods [13–15]. Compared to image-based methods that use the 3DPW dataset for training, our ARTS outperforms the state-of-the-art methods in the metrics of MPJPE, MPVPE, and Accel and competitive performance in PA-MPJPE. Due to the superior cross-dataset generalization ability of our model, ARTS without 3DPW dataset for training also achieves competitive performance. These results demonstrate that our ARTS surpasses existing image-based methods in both per-frame accuracy and temporal consistency for 3D human mesh recovery.

Additional Ablation Study. We conduct more ablation experiments to investigate the effectiveness of our temporal averaged bone length and temporal averaged human motion padding. As shown in Table 2, The ‘Frame Bone Length’ refers to using only the bone length of the mid-frame skeleton, which is not robust to errors in skeleton estimation. The ‘Zero Motion Padding’ refers to using zero vector to pad the motion sequence rather than the averaged human motion among video frames, which lacks overall human motion information. These two strategies show a decrease in per-frame accuracy and temporal consistency.

4 DETAILS ABOUT LOSS FUNCTIONS

For the training of 3D pose estimation, we use the L1 3D joint loss to supervise 3D skeletons of all frames, which is calculated as follows:

$$\mathcal{L}_{joint} = \sum_{t=1}^T \|S_t^{3D} - S_{gt}^{3D}\|_1. \quad (7)$$

For the training of the whole network, we use the following four loss functions.

Mesh Loss. We calculate the L1 loss between the predicted 3D mesh vertices $M \in \mathbb{R}^{V \times 3}$ and the ground truth 3D mesh vertices $M_{gt} \in \mathbb{R}^{V \times 3}$, where $V = 6890$ represents the number of SMPL mesh vertices [3]. The mesh loss is calculated as:

$$\mathcal{L}_{mesh} = \frac{1}{V} \sum_{i=1}^V \|M - M_{gt}\|_1. \quad (8)$$

Joint Loss. When training the whole network, we use the predicted human mesh M to regress joints through the regression matrix W . Then, we use the regressed joints of the mid-frame to calculate the joint loss:

$$\mathcal{L}_{joint} = \|WM - S_{gt}^{3D}\|_1. \quad (9)$$

Shape Loss. This loss is used to supervise the SMPL human shape. We calculate the L2 loss between the refined SMPL shape parameters $\beta_{refined}$ and the ground truth SMPL shape parameters β_{gt} of the mid-frame, which can be expressed as:

$$\mathcal{L}_{shape} = \|\beta_{refined} - \beta_{gt}\|_2. \quad (10)$$

Pose Loss. This loss is used to supervise the SMPL human pose. The pose loss is calculated by the refined SMPL pose parameters $\theta_{refined}$ and the ground truth SMPL shape parameters θ_{gt} of the mid-frame, which is calculated as:

$$\mathcal{L}_{pose} = \|\theta_{refined} - \theta_{gt}\|_2. \quad (11)$$

Given the four loss functions, the final loss is calculated as:

$$\mathcal{L} = \lambda_m \mathcal{L}_{mesh} + \lambda_j \mathcal{L}_{joint} + \lambda_p \mathcal{L}_{pose} + \lambda_s \mathcal{L}_{shape}, \quad (12)$$

where $\lambda_m = 1$, $\lambda_j = 1$, $\lambda_p = 0.06$, and $\lambda_s = 0.06$ in the experiments.

5 ADDITIONAL VISUALIZATION RESULTS

Additional Qualitative Comparison. Figure 2 shows additional qualitative comparison among the previous state-of-the-art video-methods GLoT [14], PMCE [15] and our ARTS on the challenging 3DPW dataset. Since PMCE does not use human modal prior (e.g., SMPL), it often generates unreasonable human pose and shape. Due to the effective utilization of 3D skeletons, our ARTS can produce more accurate and temporal consistent human mesh, especially in fast motions and severe occlusions.

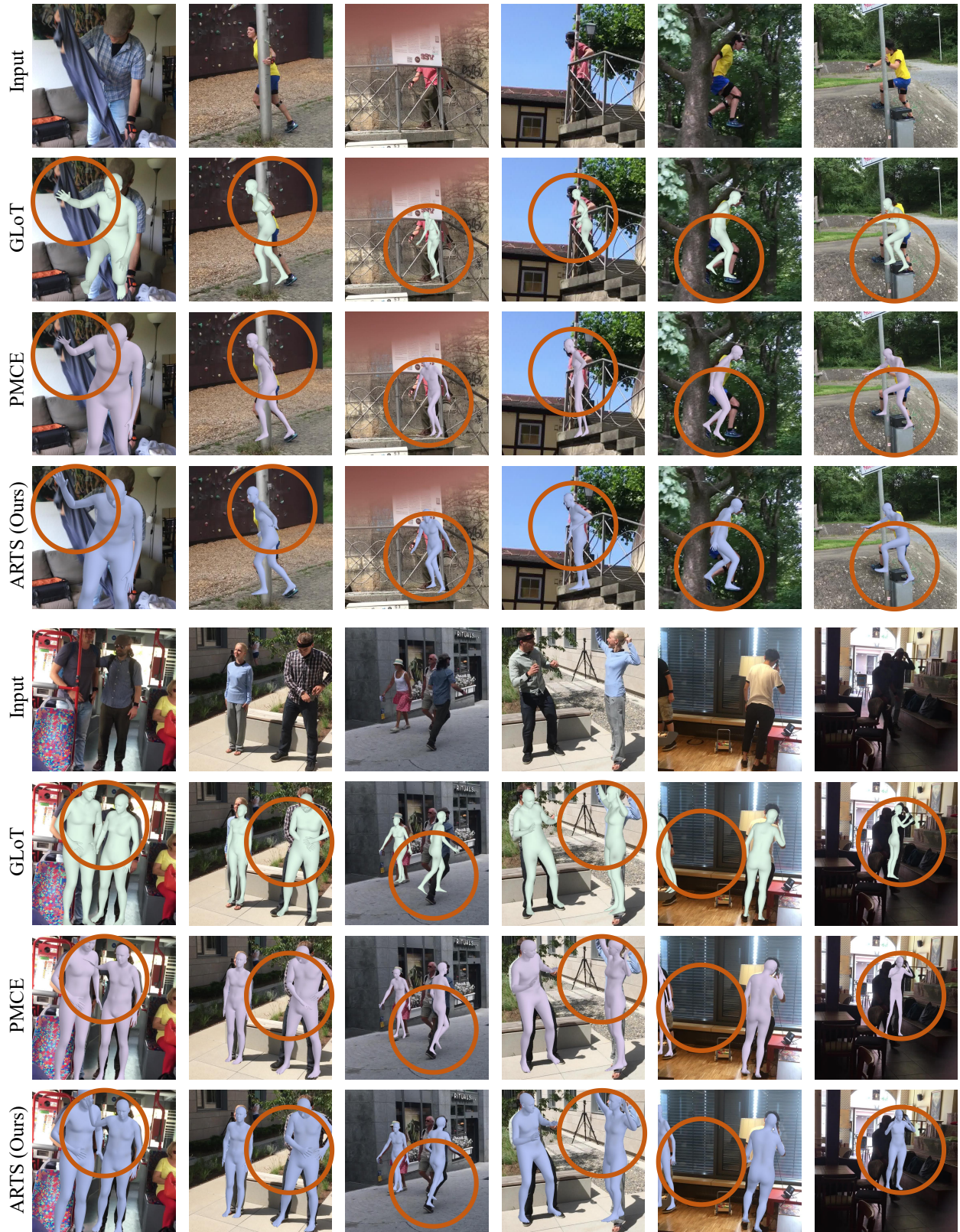


Figure 2: Additional qualitative comparison among GLoT [14] (green mesh), PMCE [15] (pink mesh) and our ARTS (blue mesh).

REFERENCES

- [1] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7103–7112, 2018.
- [2] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:38571–38584, 2022.
- [3] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866, 2023.
- [4] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11179–11188, 2021.
- [5] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 342–359. Springer, 2022.
- [6] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 590–606, 2022.
- [7] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15085–15099, 2023.
- [8] Juze Zhang, Ye Shi, Yuexin Ma, Lan Xu, Jingyi Yu, and Jingya Wang. Ikol: Inverse kinematics optimization layer for 3d human pose and shape estimation via gauss-newton differentiation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3454–3462, 2023.
- [9] Zihao Huang, Min Shi, Chengxin Liu, Ke Xian, and Zhiguo Cao. Simhmr: A simple query-based framework for parameterized human mesh reconstruction. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, pages 6918–6927, 2023.
- [10] Siyuan Bian, Jiefeng Li, Jiasheng Tang, and Cewu Lu. Shapeboost: Boosting human shape estimation with part-based parameterization and clothing-preserving augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 828–836, 2024.
- [11] Yixuan Zhu, Ao Li, Yansong Tang, Wenliang Zhao, Jie Zhou, and Jiwen Lu. Dpmesh: Exploiting diffusion prior for occluded human mesh recovery. *arXiv preprint arXiv:2404.01424*, 2024.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [13] Peng Wu, Xiankai Lu, Jianbing Shen, and Yilong Yin. Clip fusion with bi-level optimization for human mesh reconstruction from monocular videos. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, pages 105–115, 2023.
- [14] Xiaolong Shen, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Global-to-local modeling for video-based 3D human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8887–8896, 2023.
- [15] Yingxuan You, Hong Liu, Ti Wang, Wenhao Li, Runwei Ding, and Xia Li. Co-evolution of pose and mesh for 3D human body estimation from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14963–14973, 2023.