

Monocular Models are Strong Learners for Multi-View Human Mesh Recovery

Haoyu Xie¹, Shengkai Xu¹, Cheng Guo¹, Muhammad Usama Saleem¹, Wenhan Wu¹, Chen Chen², Ahmed Helmy¹, Pu Wang¹, and Hongfei Xue^{1*}

¹ University of North Carolina at Charlotte, Charlotte, NC 28223, USA
{hxie3,sxu7,cguo3,msaleem2,wwu25,ahmed.helmy,
pu.wang,hongfei.xue}@charlotte.edu

² University of Central Florida, Orlando, FL 32816, USA
chen.chen@ucf.edu

Abstract. Multi-view human mesh recovery (HMR) is broadly deployed in diverse domains where high accuracy and strong generalization are essential. Existing approaches can be broadly grouped into geometry-based and learning-based methods. However, geometry-based methods (e.g., triangulation) rely on cumbersome camera calibration, while learning-based approaches often generalize poorly to unseen camera configurations due to the lack of multi-view training data, limiting their performance in real-world scenarios. To enable calibration-free reconstruction that generalizes to arbitrary camera setups, we propose a training-free framework that leverages pretrained single-view HMR models as strong priors, eliminating the need for multi-view training data. Our method first constructs a robust and consistent multi-view initialization from single-view predictions, and then refines it via test-time optimization guided by multi-view consistency and anatomical constraints. Extensive experiments demonstrate state-of-the-art performance on standard benchmarks, surpassing multi-view models trained with explicit multi-view supervision.

Keywords: Human Mesh Recovery · Multi-view · Test-time Adaptation

1 Introduction

Multi-view human mesh recovery (HMR), which reconstructs accurate 3D human pose and shape from synchronized camera views, is a fundamental capability for high-fidelity human-centric perception. Compared with single-view methods [7, 8, 23, 39, 41] that inherently suffer from depth ambiguity, occlusions, and viewpoint-dependent uncertainty, multi-view approaches greatly improve reliability and geometric accuracy by leveraging complementary spatial information across cameras. At the same time, they offer a practical, accurate, and marker-less alternative to traditional marker-based motion capture systems [36–38, 44].

* Corresponding author.

These advantages make multi-view human mesh recovery essential for downstream applications such as fine-grained behavior understanding [40, 45], human digitalization [4, 10, 48], and immersive AR/VR experiences [32, 35, 55].

Early multi-view HMR methods [5, 11, 24, 43] have demonstrated the benefit of aggregating information from multiple viewpoints to improve reconstruction accuracy. Volumetric approaches [5, 43] typically fuse multi-view 2D features into a unified 3D feature volume via triangulation, explicitly leveraging geometric constraints during fusion. This design enables strong generalization across different camera configurations. However, this camera calibration-requiring design, as well as the inherent quantization errors introduced by discrete volumetric representations, limits their scalability in practical settings. To further improve reconstruction accuracy, recent methods [15, 22] adopt end-to-end learning frameworks that fuse multi-view features, achieving strong performance in modeling human pose and shape. However, existing multi-view datasets provide only a narrow range of camera configurations [12, 31]. As a result, the learned fusion modules inevitably memorize dataset-specific geometric priors tied to those limited camera settings.

This baked-in implicit bias severely restricts generalization: although these models perform well under training conditions, their accuracy drops sharply when evaluated on unseen camera settings. A representative example is U-HMR [22], whose feed-forward architecture aggregates image features to directly regress the 3D mesh. The model tends to memorize the 2D-to-3D correspondences observed during training, leading to substantial degradation under novel camera configurations, as demonstrated in [30]. A similar issue arises in MVP [50], where [25] shows that its performance collapses to near-zero when tested on new camera settings, indicating that it encodes specific camera parameters rather than learning robust multi-view reasoning.

In parallel, single-view HMR methods [7, 8, 23, 39, 41] have witnessed remarkable progress. A pivotal advantage of these single-view methods is their access to massive-scale training datasets [1, 12, 26, 31, 34, 49, 51], including extensive synthetic datasets [2, 28, 29]. This data advantage enables them to learn a powerful and reliable prior of the human body with diverse camera viewpoints and settings, avoiding the camera-specific generalization issues that plague multi-view learning. Landmark works have steadily advanced the state of the art: HMR2.0 [8] established a pure transformer-based baseline, TokenHMR [7] replaced the regression head with a well-trained codebook for accurate and more plausible outputs, CLIFF [23] incorporated bounding box information to improve camera-space estimation, and CameraHMR [39] integrated perspective camera model to mitigate projection error. However, these single-view solutions are inherently constrained by the fundamental ambiguities of monocular reconstruction, namely depth uncertainty and occlusion.

At this point, pure learning-based multi-view fusion modules lack generalization due to the scarcity of camera settings in the training data, but single-view methods provide a powerful and generalizable human mesh prior learned from large-scale data. This leads us to ask the following key question: *Is it possible*

to construct a multi-view HMR system that relies solely on single-view training, yet delivers high accuracy and robustness across arbitrary camera setups?

To address this question, we introduce a framework that injects single-view priors into multi-view reconstruction via iterative optimization. Our approach is built on two key insights: (1) modern single-view HMR models provide sufficiently strong and generalizable priors to serve as reliable foundations for 3D reconstruction, and (2) robust generalization to arbitrary camera configurations emerges only when multi-view fusion explicitly reasons about 3D geometry, rather than relying on purely data-driven models that implicitly memorize camera settings. To operationalize these principles, our framework directly initializes an optimizable “virtual view” representation from per-view estimates using pre-trained single-view models without any multi-view training, then optimizes this representation and per-view estimates via Test-time adaptation (TTA) guided by 2D anatomical landmarks, multi-view consistency, and single-view prior regularization. To conclude, our contributions are summarized as follows:

- We propose a multi-view TTA framework that iteratively refines single-view priors guided by multi-view information and 2D anatomical landmarks, enabling accurate and robust mesh recovery without requiring any multi-view training data.
- Through comprehensive ablation studies, we systematically explore key design choices of our TTA paradigm, including the selection of optimizable components, the formulation of guidance, and the hyper-parameter weights.
- Our method achieves state-of-the-art performance on benchmark datasets even compared to fully supervised multi-view approaches, demonstrating superior generalization to novel camera configurations.

2 Related Work

2.1 Single-view Human Mesh Recovery

Single-view human mesh recovery [3, 19, 33, 42, 54] has advanced rapidly due to large-scale datasets and strong parametric priors. Early work such as HMR [17] introduced the standard top-down pipeline that regresses SMPL parameters from image features via iterative refinement. CLIFF [23] improved global pose estimation by restoring lost spatial cues through bounding-box conditioning and full-image reprojection. HMR2.0 [8] replaced CNN encoders with a ViT-H/16 backbone, demonstrating the benefit of high-capacity transformers for mesh prediction. TokenHMR [7] addressed camera-model mismatch by proposing adaptive loss scaling and a discrete pose-token representation learned from motion capture data. Recent extensions such as CameraHMR [39] further refine camera modeling to reduce projection errors. Together, these methods provide powerful and highly generalizable single-view priors, but remain fundamentally limited by monocular ambiguities.

2.2 Multi-view Human Mesh Recovery

Multi-view HMR leverages complementary observations across cameras to resolve depth ambiguity and occlusion that fundamentally limit single-view approaches. Early geometric methods rely on triangulation or volumetric fusion [13,

43], lifting multi-view 2D cues into a discretized 3D space. Although explicitly encoding geometry and generalizing across camera setups, these approaches suffer from voxel quantization and are brittle under occlusion or imperfect calibration. Subsequent learning-based pipelines aim to fuse view-specific features directly. View-by-view refinement [24] propagates predictions across views but under-uses camera parameters and often accumulates cross-view inconsistencies. Feed-forward transformer architectures [15, 22] aggregate multi-view image features to regress SMPL parameters; however, their fusion remains implicit and tends to absorb dataset-specific camera priors, leading to poor generalization to unseen configurations. Recent iterative refinement methods [14, 30] revisit optimization-style updates for multi-view HMR. Pixel-aligned fusion [14] reprojects intermediate meshes to extract feedback signals, while HeatFormer [30] formulates multiview alignment as heatmap matching via a transformer-based neural optimizer. These approaches improve robustness to occlusion and camera variation, however, their performance degrades in calibration-free mode. HeatFormer relies on epipolar geometry in AdaFuse [56] module to align cross-view heatmaps; without reliable calibration, cross-view information cannot be effectively aggregated. Similarly, PaFF [14] depends on camera parameters for pixel-aligned feedback, orientation alignment, and scale recovery. Moreover, both methods require multi-view training data and may implicitly encode dataset-specific camera layouts, limiting their generalization to unseen configurations.

3 Method

3.1 Overview

Our goal is to recover a robust and consistent 3D human mesh from multiple camera views while generalizing to unseen camera configurations, without relying on multi-view training data. To achieve it, we build on pretrained single-view HMR models and refine their predictions at inference time by optimizing a small set of variables under multi-view constraints and 2D anatomical clues. However, this design raises two key questions: (1) *what to optimize*—which parts of a pretrained single-view model should be updated at test time to enable effective adaptation; and (2) *how to optimize*—which constraints should guide the updates to enforce cross-view consistency and anatomical plausibility. We answer these questions by (i) analyzing different choices of optimizable components and their initialization in the multi-view setting, and (ii) introducing a set of test-time objectives that drive refinement toward coherent solutions.

The remainder of this section is organized as follows. Sec. 3.2 briefly reviews the required preliminaries. Sec. 3.3 outlines the overall inference-time optimization pipeline. Sec. 3.4 studies the choice and initialization of optimizable components. Sec. 3.5 presents the test-time objectives used to guide refinement.

3.2 Preliminaries

SMPL Model. The SMPL model [27] represents the human body using a low-dimensional parametric formulation. It is controlled by pose parameters $\theta \in \mathbb{R}^{72}$,

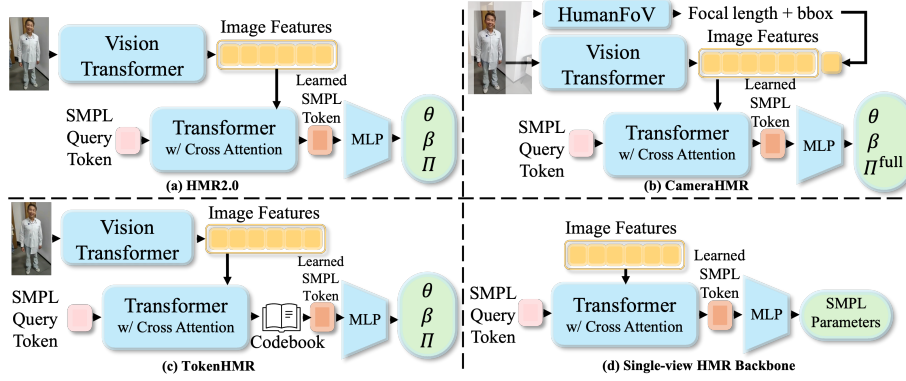


Fig. 1: The model architectures of (a) HMR2.0, (b) CameraHMR, (c) TokenHMR, and (d) a summary of these generic single-view transformer-based HMR models. The optimizable components can be selected from three representative groups: tokens (pink/orange), model parameters (blue), or explicit SMPL parameters (green).

encoding the relative 3D rotations of 24 joints, and shape parameters $\beta \in \mathbb{R}^{10}$, capturing person-specific body shape variations. Given (θ, β) , SMPL outputs a posed human mesh $V = M(\theta, \beta) \in \mathbb{R}^{6890 \times 3}$ using linear blend skinning. The corresponding 3D joint coordinates $J_{3D} \in \mathbb{R}^{K \times 3}$ are obtained by applying a fixed linear joint regressor to the mesh. **In our framework**, SMPL model is used for generating human mesh from SMPL parameters.

Single-view HMR Backbones. Recent transformer-based single-view HMR models directly regress the SMPL parameters (θ, β) and camera coefficients Π from a single RGB image. HMR2.0 [8] (Fig. 1(a)) establishes a strong baseline by adopting a Vision Transformer (ViT [6]) encoder to extract image features and a transformer decoder that operates on learnable SMPL tokens, which are finally mapped to (θ, β) via an SMPL regression head. Building upon this architecture, CameraHMR [39] (Fig. 1(b)) improves reconstruction by explicitly modeling the perspective camera. It predicts the camera field-of-view and introduces camera-aware tokens into the transformer to enable more accurate projection. TokenHMR [7] (Fig. 1(c)) explores a complementary direction by replacing continuous pose regression with a tokenized pose representation. It predicts pose tokens decoded via a learned codebook, providing stronger pose priors and more robust reconstruction. Despite these design differences, these approaches share a common paradigm that we refer to as the single-view HMR backbone (Fig. 1(d)). Such models typically consist of a ViT encoder that extracts image features and a transformer-based learning structure that updates learnable SMPL tokens through attention with the image features. The resulting token representation is then mapped to the SMPL parameters (θ, β) and camera coefficients Π . **In our framework**, these pretrained single-view HMR backbones are leveraged as strong priors to enable generalizable multi-view reconstruction without requiring any multi-view training data. These models provide high-quality

initial estimates and impose powerful single-view regularization during test-time optimization, encouraging per-view poses to remain consistent with the image-conditioned latent space learned from large-scale single-view datasets.

2D Keypoint Detector. A pretrained and frozen 2D keypoint detector can be employed to extract reliable 2D image-space cues for each view. Depending on the detector’s training objective, it can provide sparse joint locations, dense correspondences, or mesh-aligned keypoints directly from the input image. These detections supply strong geometric evidence that is independent of the single-view HMR backbone. **In our framework**, these detected 2D keypoints serve as robust constraints on the projection of the 3D joints during test-time optimization, effectively guiding the refinement of pose and shape parameters and improving reconstruction accuracy.

3.3 Proposed Paradigm

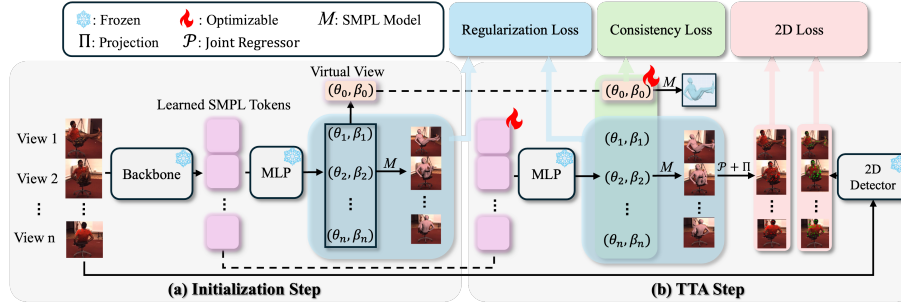


Fig. 2: Test-time adaptation (TTA) framework for multi-view human mesh recovery. The pipeline is composed of two stages: (a) aggregating single-view SMPL estimates into a set of SMPL parameters initialization termed *virtual view*; (b) Test-time adaptation refines single-view optimizable token and virtual view SMPL parameters using 2D Loss, multi-view consistency Loss, and regularization loss.

To generalize to arbitrary camera configurations, we leverage strong priors from pretrained single-view HMR models, then refine them at inference time, therefore eliminating the multi-view training. Based on this principle, we propose a model-agnostic TTA paradigm for multi-view human mesh recovery. Our core approach treats single-view predictions as an initialization, iteratively refining them through cross-view consistency and 2D anatomical cues, while regularization terms prevent the results from drifting excessively from the reliable single-view priors. In Fig. 2(a), a frozen single-view HMR backbone first produces per-view SMPL estimates. In addition, we introduce an extra set of optimizable SMPL parameters, termed the *virtual view*. The virtual view is initialized from the per-view SMPL predictions. During TTA, the virtual view parameters are further refined using gradients aggregated from all views. After optimization, the virtual view represents the final reconstruction that summarizes the information from all input views. Conceptually, the virtual view plays a role similar to aggregating multi-view predictions (e.g., averaging per-view results), but instead

of directly fusing predictions, it is optimized through gradients from all views. In Fig. 2(b), TTA jointly refines the learned SMPL tokens and the virtual-view parameters through 2D observations, multi-view consistency constraints, and regularization terms.

3.4 Optimizable Component

The choice of the optimizable component Θ is crucial for effective test-time adaptation, as the selected component must balance expressiveness, computational efficiency, and alignment with the adaptation objectives.

Single-view Backbone Optimizable Component. As illustrated in Fig. 1, there are multiple candidate components within transformer-based single-view HMR models that can be optimized during test-time adaptation within our proposed paradigm, which are broadly categorized into three groups: model parameters (blue), latent tokens (pink and orange), and explicit SMPL parameters (green). **Model parameters** encompass the weights of the transformer backbone and decoder head, which encode rich priors learned from large-scale single-view datasets. However, these weights are extremely high-dimensional, making test-time optimization computationally prohibitive and potentially destructive to the pretrained priors. **Explicit SMPL parameters** are low-dimensional (72 for pose θ and 10 for shape β), making them computationally efficient to optimize. However, optimizing only these parameters fails to exploit the image-conditioned latent structure learned by single-view HMR models. Moreover, the optimization landscape with respect to the SMPL parameters is highly non-linear, which makes the process prone to getting trapped in poor local minima. As a result, the optimization can easily overfit noisy 2D detections or drift toward view-inconsistent solutions. **Tokens** provide a compact representation for modeling human pose and shape within transformer-based HMR architectures. Within this category, two common variants can be considered. SMPL query tokens (pink in Fig. 1) serve as the initial input tokens to the transformer, lacking sufficiently abstracted and task-specific features to fully support stable multi-view adaptation. In contrast, learned SMPL tokens (orange in Fig. 1) capture richer semantic and image-conditioned representations that are directly instrumental for pose and shape prediction. During optimization of the learned SMPL tokens, gradients are propagated only through the final MLP layers (implemented as a single linear projection) without back-propagating through the transformer backbone. **In our framework**, we therefore adopt learned SMPL tokens as the optimization variables within single-view backbone. This lightweight design avoids full forward and backward passes through the entire network, enabling lightweight TTA while preserving the pretrained backbone priors. Ablation comparisons of different optimization components are provided in Sec. 4.4.

Virtual-View Optimizable Component. To obtain a unified reconstruction across views, we introduce a set of optimizable SMPL parameters, termed the *virtual view*, which is initialized from per-view predictions and subsequently updated during TTA. We initialize the virtual-view pose parameters using a weighted averaging strategy over the per-view predictions. Specifically, we com-

pute per-joint statistics across views and apply a $1-\delta$ filtering mechanism to suppress unreliable joint rotations. For each joint rotation k , represented using the continuous 6D parameterization [57], we retain the reliable set $\mathcal{S}^{(k)}$:

$$\mathcal{S}^{(k)} = \left\{ \theta_i^{(k)} \mid d_i^{(k)} \leq \text{Std}(d_1^{(k)}, \dots, d_N^{(k)}) \right\}, s.t. \quad d_i^{(k)} = \left\| \theta_i^{(k)} - \frac{1}{N} \sum_{i=1}^N \theta_i^{(k)} \right\|_2, \quad (1)$$

where $i \in \{1, \dots, N\}$ is the camera view index, and $\theta_i^{(k)}$ denotes the rotation of joint k from view i . We define $i = 0$ as the virtual view. The virtual-view pose parameters $\theta_0^{(k)}$ are initialized as the mean of the retained set $\mathcal{S}^{(k)}$. For the global orientation (i.e., the root joint rotation $\theta_i^{(0)}$), two settings are considered. When cameras are calibrated, we first transform the per-view global orientations into the world coordinate frame using the known camera extrinsics, and then apply the weighted averaging described above. When cameras are uncalibrated (calibration-free mode), global orientations remain view-specific and are not included in the weighted averaging procedure, since they cannot be expressed in a shared coordinate frame. In this case, only the global orientation is view-specific, while all remaining pose rotations are weighted averaged. Shape parameters are assumed to be view-consistent and initialized as $\beta_0 = \frac{1}{N} \sum_{i=1}^N \beta_i$. We denote the virtual-view optimizable SMPL parameters as $\Theta_0 = (\theta_0, \beta_0)$. We empirically observe that this reliability-aware weighted initialization improves stability and final performance compared to simply averaging or canonical pose initialization. Further analysis of different initialization strategies is provided in Sec. 4.4. In our framework, we jointly optimize the learned SMPL tokens within single-view backbone and the virtual-view SMPL parameters.

3.5 Test-time Adaptation

We formalize the TTA paradigm as follows:

$$\Theta_{i,t+1} = \Theta_{i,t} - \eta \nabla_{\Theta_{i,t}} (\mathcal{L}_{2D} + \lambda \mathcal{L}_{\text{con}} + \gamma \mathcal{L}_{\text{reg}}) \quad (2)$$

at each optimization step t during inference, we adaptively update the optimizable component $\Theta_{i,t}$ under the guidance of three losses: the 2D reprojection loss $\mathcal{L}_{2D}$, the cross-view consistency loss $\mathcal{L}_{\text{con}}$, and the regularization loss $\mathcal{L}_{\text{reg}}$. The weights λ, γ balance the contributions of these losses, and η denotes the learning rate. We will alternatively update the single-view backbone optimizable component (learned SMPL token) and the virtual view component (SMPL parameters) with their respective losses. For simplicity, the settings of the optimization step, the loss weights, and other hyperparameters are omitted from the loss notation; detailed settings are provided in the Supplementary.

2D Reprojection Loss. To provide view-specific supervision during test-time adaptation, we leverage a pretrained DensePose-based anatomical landmark detector [9], which supplies pseudo-ground truth annotations $\hat{P}_i$ for each view. These annotations include 44 joint keypoints (25 OpenPose joints and 19 additional joints) and 35 anatomical surface landmarks, each associated with a

confidence score $D_i^{(k)}$. We incorporate anatomical landmarks because, in contrast to standard joint keypoints, they encode not only spatial positions but also fine-grained joint twist rotation cues [21]. For example, a wrist twist alters the appearance of the hand surface even though the 3D wrist joint location barely changes. Such rotational effects produce significant 2D displacements in the surface landmarks, enabling the supervision to capture subtle pose variations that joint-only signals fail to represent. To mitigate detector noise, we retain only 2D observations with confidence greater than 0.9. For each camera view i , the current optimizable component Θ_i (i.e., the learned SMPL token) is mapped to a 3D mesh $\mathcal{M}_i$ through the MLP and SMPL wrapper, denoted as $\mathcal{M}_i = \mathcal{R}_i(\Theta_i)$. We then regress 44 joint locations from $\mathcal{M}_i$ using a pretrained keypoint regressor $\mathcal{P}(\cdot)$ [8], and project them to the image plane: $P_i^r = \Pi_i(\mathcal{P}(\mathcal{M}_i))$, where Π_i denotes the camera projection function. The 35 anatomical landmarks P_i^a are obtained directly by indexing the corresponding surface vertices from $\mathcal{M}_i$ via a predefined index list. The 2D reprojection loss for view i is defined as:

$$\mathcal{L}_{2d} = \|P_i^r - \hat{P}_i^r\|_2 + \|P_i^a - \hat{P}_i^a\|_2$$

Cross-view Consistency Loss. Cross-view consistency loss is added to enforce agreement of human mesh and geometry between pairs of views (i, j) , which is calculated by:

$$\mathcal{L}_{\text{con}} = \sum_{i \neq j}^N \|\theta_i - \theta_j\|_2 + \|\beta_i - \beta_j\|_2 + \|\mathcal{M}_i - \mathcal{M}_j\|_2$$

The above formulation is applied differently depending on the optimized component. When updating the virtual-view component, the loss is computed between the virtual view ($i = 0$) and each camera view ($j \in \{1, \dots, N\}$), and gradients are applied only to the virtual-view parameters. When optimizing the single-view backbone components, the summation is restricted to camera views ($i, j \in \{1, \dots, N\}$), enforcing pairwise agreement among them. In calibration-free mode, the global-orientation cannot be applied, as per-view orientations are not expressed in a common frame. The quadratic complexity of $\mathcal{O}(N^2)$ with respect to the number of views can be reduced to $\mathcal{O}(N)$ by adopting a star-structured consistency graph with negligible performance difference. Further details are provided in the supplementary.

Regularization Loss. A regularization loss is also added to prevent large deviations from the initial single-view predictions:

$$\mathcal{L}_{\text{reg}} = \sum_{i=1}^N \|\theta_i - \theta_{i,0}\|_2 + \|\beta_i - \beta_{i,0}\|_2 + \|\mathcal{M}_i - \mathcal{M}_{i,0}\|_2$$

where $\theta_{i,0}$, $\beta_{i,0}$, and $\mathcal{M}_{i,0}$ denote the parameters and mesh of view i at iteration $t = 0$, i.e., the initial predictions obtained before test-time adaptation.

4 Experiments

4.1 Experiment Configuration

Pretrained Single-view Backbones. HMR2.0b is trained on Human3.6M [12], MPI-INF-3DHP [31], COCO [26], and MPII [1], together with pseudo-ground-truth SMPL fits on AVA [34], InstaVariety [18], and AI Challenger [51]. TokenHMR trains its tokenizer with AMASS [28] and MOYO [49], and further trains the HMR regressor with the HMR2.0b training setting plus BEDLAM [2]. CameraHMR uses improved 4DHumans pseudo-ground truth with perspective camera modeling and dense surface keypoints, and is trained with improved 4DHumans data together with BEDLAM and AGORA, excluding AVA.

2D Detectors. We choose a pretrained 2D detector, Synthpose [9]. It’s first pretrained on the original COCO, COCO-WholeBody dataset [16,52], then finetuned on the 3DPW and BEDLAM training dataset. We then adapt the network’s head to output 35 anatomical points and 44 skeleton points, and finetune it on the Human3.6M and MPI-INF-3DHP training dataset. Following previous multi-view work [24, 43], we train the detector on subjects 1,5,6,7,8 and test on subjects 9, 11 for Human3.6M, and train on subjects 1 to 7 and test on subject 8 for MPI-INF-3DHP. The reason we fine-tune is that the SMPL parameters in Human3.6M and MPI-INF-3DHP datasets are pseudo ground truths generated by NeuralAnnot [33], which exhibits systematic misalignment between the SMPL mesh and the real subjects. To ensure consistency between the detected 2D keypoints and the pseudo-GT SMPL annotations used during TTA, we fine-tune the detector on these datasets.

4.2 Evaluation Metrics

Following previous works [24, 43], we evaluate 3D human mesh recovery on Human3.6M using MPJPE (Mean Per Joint Position Error), PA-MPJPE (Procrustes Analysis Mean Per Joint Position Error), and MPVPE (Mean Per Vertex Position Error). MPJPE/MPVPE measures the Euclidean distance between predicted and ground truth joints/vertices after aligning the pelvis, and PA means procrustes aligned. We evaluate on MPI-INF-3DHP using MPJPE, PCK (Percentage of Correct Keypoints), and AUC (Area Under the Curve). PCK means the percentage of keypoints having less than or equal to a predetermined distance from the ground truth. AUC means the area under the PCK curve.

We evaluate our method using three representative single-view backbones: CameraHMR [39], HMR2.0 [8], and TokenHMR [7]. Unless otherwise stated, we follow a *Standard* setup, which corresponds to our complete multi-view test-time adaptation pipeline. This standard setup includes: (1) the 1-delta filtered virtual-view fusion for producing a robust multi-view initialization, (2) 2D clues from both keypoints and anatomical landmarks, (3) multi-view consistency loss together with a regularizer that anchors the optimization to the initial predictions, and (4) using the learned token as the evaluation component to be optimized during TTA. All reported results default to this standard setup except results in the ablation study.

Table 1: Comparison on the Human3.6M dataset. Methods are grouped into single-view, multi-view calibration-free, and multi-view calibration-requiring approaches. **Bold** and underline indicate the best and second-best results within each group, respectively.

	Method	MPJPE ↓	PA-MPJPE ↓
Single Camera	HMR2.0(a) [8]	44.8	33.6
	HMR2.0(b) [8]	50.0	32.4
	HMR2.0+scoreHMR-a [47]	47.9	28.4
	HMR2.0+scoreHMR-b [47]	44.7	29.0
	PostureHMR [46]	44.5	31.0
Calibration-free	Shape Aware [24]	79.9	45.1
	ProHMR [20]	62.2	34.5
	Yu et al [53]	-	33.0
	PaFF [14]	44.8	28.2
	HeatFormer [30]	42.5	25.8
	U-HMR [22]	31.0	<u>22.8</u>
	Ours (TokenHMR)	44.3	23.9
	Ours (HMR2.0)	45.2	23.0
	Ours (CameraHMR)	<u>32.7</u>	22.0
w. Calibration	MuVS [11]	58.2	47.1
	MV-SPIN [43]	49.8	35.4
	PaFF [14]	33.0	26.9
	HeatFormer [30]	<u>29.5</u>	22.4
	Ours (TokenHMR)	32.6	23.2
	Ours (HMR2.0)	31.6	<u>21.5</u>
	Ours (CameraHMR)	26.9	20.6

4.3 Comparison with State-of-the-art Methods

Notably, our method is trained solely on single-view data without any multi-view supervision, all the following experiments consistently demonstrate its strong generalization to multi-view camera settings.

Human3.6M. Table 1 reports quantitative results on Human3.6M under the standard protocol. In calibration setting, our method with CameraHMR backbone achieves 26.9 MPJPE and 20.6 PA-MPJPE, substantially outperforming the traditional multi-view SMPL fitting method MuVS, the learning-based method MV-SPIN, the neural optimization approach PaFF, and the previous state of the art HeatFormer (29.5 MPJPE / 22.4 PA-MPJPE). Notably, our calibration-free setting achieves substantially lower errors (32.7 MPJPE / 22.0 PA-MPJPE) compared to HeatFormer (42.5 MPJPE / 25.8 PA-MPJPE), highlighting the effectiveness of leveraging strong single-view priors for multi-view refinement without requiring camera calibration. Among the backbones, CameraHMR backbone exhibits particularly strong robustness in the calibration-free setting. We attribute this to its perspective camera formulation and explicit focal length estimation. These results further validate that a strong single-view prior, combined with TTA, enables effective multi-view reconstruction even without camera calibration.

Table 2: Comparison results on MPI-INF-3DHP dataset with multi-view methods. **Bold** indicates the best result, Underline indicates the second-best result.

Method	MPJPE ↓	PCK ↑	AUC ↑
PaFF [14]	48.4	98.6	67.3
U-HMR [22]	<u>39.7</u>	99.4	74.0
HeatFormer [30]	39.8	99.5	72.8
Ours (TokenHMR)	45.2	99.9	79.6
Ours (HMR2.0)	40.3	<u>99.6</u>	<u>80.7</u>
Ours (CameraHMR)	39.0	99.9	83.8

Table 3: Comparison results on MPI-INF-3DHP dataset with cross camera settings.

Method	MPJPE ↓	PCK ↑	AUC ↑
HeatFormer	45.74	99.15	68.93
Ours(CameraHMR)	43.71	99.44	80.62

MPI-INF-3DHP. The comparison on the MPI-INF-3DHP dataset, as shown in Table 2, also highlights our work’s performance and generalization. The CameraHMR backbone outperforms previous methods, demonstrating the strong performance and generalizability of our approach.

Cross-camera Results. Table 3 evaluates cross-camera generalization by testing on MPI-INF-3DHP camera set (1, 4, 5, 6), which differs from the original set (0, 2, 7, 8). As shown in Table 3, HeatFormer [30] degrades notably under this camera shift, achieving 45.74 MPJPE. Our method, however, reaches 43.71 MPJPE, 99.44 PCK, and 80.62 AUC, outperforming HeatFormer across all metrics. This demonstrates that our geometry-driven TTA adapts effectively to unseen camera settings, offering strong cross-camera robustness without requiring multi-view training data.

4.4 Ablation Study

Impacts of Virtual-View Initialization and TTA. Table 4 shows that the best performance is achieved when both components are enabled. Most of the improvement comes from TTA, which substantially refines the initial predictions. Without TTA, the initialized virtual view provides only limited gains, indicating that initialization alone cannot resolve multi-view ambiguities but serves as a stable starting point for optimization.

Impact of TTA Steps. Table 5 reports the impact of TTA steps. CameraHMR continues to benefit from additional optimization steps, with the best results obtained at 200 steps. For the HMR2.0, the best performance is reached at 150 steps and remains very stable with negligible variation as the number of steps increases. TokenHMR shows a similar trend, steadily improving as the number of TTA steps increases and achieving its best performance at 200 steps, with only a marginal degradation at 250 steps. Considering both performance and generalization across different backbones, we adopt 200 TTA steps as the default setting in all experiments.

Table 4: Ablation study on the effects of Virtual View design and TTA.

Components		CameraHMR		HMR2.0		TokenHMR	
Virtual View	TTA	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
✗	✓	<u>27.8</u>	<u>22.7</u>	<u>35.1</u>	24.1	43.9	27.5
✓	✗	44.2	29.9	44.4	<u>24.0</u>	<u>37.1</u>	<u>25.7</u>
✓	✓	26.9	20.6	31.6	21.5	32.6	23.2

Impacts of Virtual-View Initialization Strategies. Table 6 compares different initialization strategies for the virtual-view component. “No virtual view” directly uses the averaged per-view estimates after TTA as the final output. “T-pose” initializes the virtual view with a canonical pose, while “Averaged” simply averages per-view SMPL parameters. “Weighted” follows the weighted-average strategy in Sec. 3.4. Across all backbones, introducing a virtual view consistently improves performance, indicating that a unified representation benefits multi-view optimization. Compared to CameraHMR, HMR2.0 and TokenHMR are more sensitive to initialization quality.

Table 5: Ablation study on TTA steps.

Steps	CameraHMR		HMR2.0		TokenHMR	
	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
50	35.6	25.9	35.4	<u>22.0</u>	35.1	24.3
100	29.9	22.3	32.4	21.5	33.5	23.6
150	28.0	21.2	31.6	21.5	32.8	<u>23.3</u>
200	26.9	20.6	31.6	21.5	32.6	23.2
250	<u>27.4</u>	<u>21.0</u>	<u>31.7</u>	21.5	<u>32.7</u>	<u>23.3</u>

Impacts of TTA Loss Components. We isolate the effect of different guidance signals used during TTA. We consider five types of settings: (1) only 2D anatomical landmarks, (2) only 2D keypoints, (3) their combination, (4) an additional multi-view consistency, and (5) The standard setting further includes a regularization term. Reported in Table 7, using either anatomical landmarks or keypoints alone provides moderate improvements, while combining them yields stronger 2D supervision. The largest performance gain comes from introducing the multi-view consistency term, which effectively exploits geometric relationships across views. Adding the regularization term delivers the best and most stable results across all three backbones.

Impacts of Single-view Backbone Optimizable Component. As shown in Tab. 8, optimizing the learned SMPL tokens achieves the best performance across all backbones, yielding the lowest MPJPE and PA-MPJPE. We hypothesize that these tokens capture rich image-conditioned representations that are directly used for pose and shape prediction. In contrast, optimizing the SMPL query tokens or decoder parameters is less effective, as the former lacks sufficient task-specific semantics and the latter is less amenable to refinement from a single test

Table 6: Ablation study on virtual view initialization strategy.

Strategy	CameraHMR		HMR2.0		TokenHMR	
	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
No Virtual View	27.8	22.7	35.1	24.1	43.9	27.5
T-pose	27.2	21.2	32.5	22.4	33.8	23.6
Averaged	<u>27.1</u>	<u>20.7</u>	<u>32.1</u>	<u>21.7</u>	<u>33.5</u>	<u>23.3</u>
Weighted	26.9	20.6	31.6	21.5	32.6	23.2

Table 7: Ablation study on TTA loss components.

Guidance	CameraHMR		HMR2.0		TokenHMR	
	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
anatomical (ana)	39.9	29.4	35.2	24.6	33.5	24.5
keypoint (kp)	39.6	28.5	35.0	24.0	33.2	23.7
kp+ana	36.8	27.2	32.3	22.5	33.0	23.6
kp+ana+consistency (con)	<u>27.3</u>	<u>20.9</u>	<u>32.2</u>	<u>21.6</u>	<u>32.9</u>	<u>23.4</u>
kp+ana+con+regularizer	26.9	20.6	31.6	21.5	32.6	23.2

Table 8: Ablation study on single-view backbone optimizable component.

Component	CameraHMR		HMR2.0		TokenHMR	
	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
SMPL Parameters	58.2	32.2	53.0	31.9	48.4	31.1
Decoder	30.8	22.8	37.2	23.3	36.9	27.5
SMPL Query Tokens	<u>29.7</u>	<u>20.9</u>	<u>36.0</u>	<u>23.0</u>	<u>35.2</u>	<u>25.7</u>
Learned SMPL Tokens	26.9	20.6	31.6	21.5	32.6	23.2

Table 9: Ablation study on Virtual-view Optimizable Component.

Component	CameraHMR		HMR2.0		TokenHMR	
	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
Pose	31.6	21.9	35.2	24.6	34.6	23.3
Pose+Orientation	<u>29.0</u>	<u>21.4</u>	<u>33.4</u>	<u>22.0</u>	<u>34.4</u>	<u>23.3</u>
Pose+Orientation+Shape	26.9	20.6	31.6	21.5	32.6	23.2

sample. Directly optimizing SMPL parameters performs the worst due to their limited dimensionality, and susceptibility to noisy 2D cues.

Impacts of Virtual-view Optimizable Component. Ablations on optimizing virtual-view components shown in Tab. 9 demonstrate that jointly adapting pose, global orientation, and shape consistently achieves the lowest MPJPE and PA-MPJPE across all backbones. Updating pose only leads to suboptimal performance due to unresolved orientation and shape mismatches, while adding orientation already yields clear gains. Incorporating shape further stabilizes cross-view alignment. While CameraHMR and HMR2.0 exhibit large gains from adding orientation and shape, TokenHMR shows smaller improvements, likely due to its discrete pose prior constraining updates.

4.5 Qualitative Result

Fig. 3 shows the predicted SMPL mesh on Human3.6M and MPI-INF-3DHP. At Step 0, the initial single-view predictions often exhibit noticeable errors. After 200 TTA steps, these errors are consistently corrected. Fig. 4 visualizes the progression of the test-time adaptation. As the TTA steps increase, the

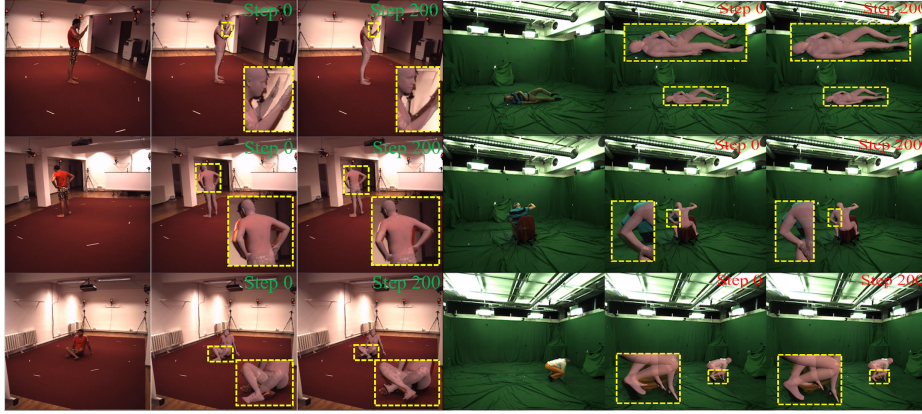


Fig. 3: Mesh alignment on the Human3.6M [12] (left) and MPI-INF-3DHP [31] (right). From left to right: input image, mesh at Step 0, and mesh at step 200.

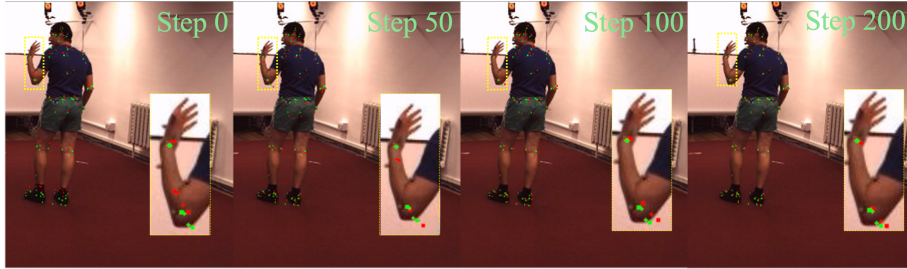


Fig. 4: TTA 2D clues alignment. Green markers denote the pseudo ground-truth from 2D detector, while the red markers represent the model’s predictions.

predicted 2D joints gradually converge toward the pseudo-GT keypoints. More qualitative results are presented in Supplementary.

5 Conclusion

In this paper, we present a training-free framework for robust and accurate multi-view human mesh recovery. By leveraging pretrained single-view HMR models as strong priors, our method introduces a virtual-view representation that is iteratively refined via test-time adaptation guided by 2D anatomical landmarks and multi-view consistency. Extensive experiments demonstrate that our approach achieves state-of-the-art performance on benchmark datasets and generalizes effectively to unseen multi-view camera configurations without requiring any multi-view training data.

Acknowledgments

This work was supported in part by the US National Science Foundation under Grants IIS-2348427 and 2008447.

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2014)
2. Black, M.J., Patel, P., Tesch, J., Yang, J.: BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 8726–8737 (Jun 2023)
3. Bogu, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: European conference on computer vision. pp. 561–578. Springer (2016)
4. Chan, K.T.: Emergence of the ‘digitalized self’ in the age of digitalization. *Computers in Human Behavior Reports* **6**, 100191 (2022)
5. Chun, S., Park, S., Chang, J.Y.: Learnable human mesh triangulation for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2850–2859 (2023)
6. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Dwivedi, S.K., Sun, Y., Patel, P., Feng, Y., Black, M.J.: Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1323–1333 (2024)
8. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4d: Reconstructing and tracking humans with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14783–14794 (2023)
9. Gozlan, Y., Falisse, A., Uhlich, S., Gatti, A., Black, M., Hicks, J., Delp, S., Chaudhari, A.: Opencapbench: A benchmark to bridge pose estimation and biomechanics. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4056–4065. IEEE (2025)
10. Hassani, H., Huang, X., Silva, E.: The human digitalisation journey: Technology first at the expense of humans? *Information* **12**(7), 267 (2021)
11. Huang, Y., Bogu, F., Lassner, C., Kanazawa, A., Gehler, P.V., Romero, J., Akhter, I., Black, M.J.: Towards accurate marker-less human shape and pose estimation over time. In: 2017 international conference on 3D vision (3DV). pp. 421–430. IEEE (2017)
12. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI* **36**(7), 1325–1339 (2014)
13. Isakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7718–7727 (2019)
14. Jia, K., Zhang, H., An, L., Liu, Y.: Delving deep into pixel alignment feature for accurate multi-view human mesh recovery. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 989–997 (2023)
15. Jiang, X., Nie, X., Wang, Z., Liu, L., Liu, S.: Multi-view human body mesh translator. arXiv preprint arXiv:2210.01886 (2022)
16. Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., Luo, P.: Whole-body human pose estimation in the wild. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)

17. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7122–7131 (2018)
18. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5614–5623 (2019)
19. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2252–2261 (2019)
20. Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11605–11614 (2021)
21. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3383–3393 (2021)
22. Li, X., Meng, M., Wu, Z., Chen, T., Yang, F., Shen, D.: Human mesh recovery from arbitrary multi-view images. arXiv preprint arXiv:2403.12434 (2024)
23. Li, Z., Liu, J., Zhang, Z., Xu, S., Yan, Y.: Cliff: Carrying location information in full frames into human pose and shape estimation. In: European Conference on Computer Vision. pp. 590–606. Springer (2022)
24. Liang, J., Lin, M.C.: Shape-aware human pose and shape reconstruction using multi-view images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4352–4362 (2019)
25. Liao, Z., Zhu, J., Wang, C., Hu, H., Waslander, S.L.: Multiple view geometry transformers for 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 708–717 (2024)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
27. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)* **34**(6), 1–16 (2015)
28. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5442–5451 (2019)
29. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: European Conference on Computer Vision (ECCV) (sep 2018)
30. Matsubara, Y., Nishino, K.: Heatformer: A neural optimizer for multiview human mesh recovery. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6415–6424 (2025)
31. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 3D Vision (3DV), 2017 Fifth International Conference on. IEEE (2017). <https://doi.org/10.1109/3dv.2017.00064>, http://gvp.mpi-inf.mpg.de/3dhp_dataset
32. Moloney, J., Spehar, B., Globa, A., Wang, R.: The affordance of virtual reality to enable the sensory representation of multi-dimensional data for immersive analytics: from experience to insight. *Journal of Big Data* **5**(1), 53 (2018)

33. Moon, G., Choi, H., Lee, K.M.: Neuralannot: Neural annotator for 3d human mesh training sets. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2299–2307 (2022)
34. Murray, N., Marchesotti, L., Perronnin, F.: Ava: A large-scale database for aesthetic visual analysis. In: *2012 IEEE conference on computer vision and pattern recognition*. pp. 2408–2415. IEEE (2012)
35. Neo, J.R.J., Won, A.S., Shepley, M.M.: Designing immersive virtual environments for human behavior research. *Frontiers in Virtual Reality* **2**, 603750 (2021)
36. Pagnon, D., Domalain, M., Reveret, L.: Pose2sim: An end-to-end workflow for 3d markerless sports kinematics—part 1: Robustness. *Sensors* (2021). <https://doi.org/10.3390/s21196530>, <https://www.mdpi.com/1424-8220/21/19/6530>
37. Pagnon, D., Domalain, M., Reveret, L.: Pose2sim: An end-to-end workflow for 3d markerless sports kinematics—part 2: Accuracy. *Sensors* (2022). <https://doi.org/10.3390/s22072712>, <https://www.mdpi.com/1424-8220/22/7/2712>
38. Pagnon, D., Domalain, M., Reveret, L.: Pose2sim: An open-source python package for multiview markerless kinematics. *Journal of Open Source Software* (2022). <https://doi.org/10.21105/joss.04362>, <https://joss.theoj.org/papers/10.21105/joss.04362>
39. Patel, P., Black, M.J.: Camerahmr: Aligning people with perspective. In: *2025 International Conference on 3D Vision (3DV)*. pp. 1562–1571. IEEE (2025)
40. Ragusa, F., Leonardi, R., Mazzamuto, M., Bonanno, C., Scavo, R., Furnari, A., Farinella, G.M.: Enigma-51: Towards a fine-grained understanding of human behavior in industrial scenarios. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 4549–4559 (2024)
41. Saleem, M.U., Pinyoanuntapong, E., Wang, P., Xue, H., Das, S., Chen, C.: Genhmr: Generative human mesh recovery. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 39, pp. 6749–6757 (2025)
42. Sengupta, A., Budvytis, I., Cipolla, R.: Humaniflow: Ancestor-conditioned normalising flows on so (3) manifolds for human pose and shape distribution estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4779–4789 (2023)
43. Shin, S., Halilaj, E.: Multi-view human pose and shape estimation using learnable volumetric aggregation. *arxiv.org*. arXiv preprint arXiv:2011.13427 (2020)
44. Shuai, Q., Fang, Q., Dong, J., Peng, S., Huang, D., et al.: Easymocap-make human motion capture easier. *Github* **1**(3), 6 (2021)
45. Singh, M., Duggirala, M., Hayatnagarkar, H., Patel, S., Balaraman, V.: Towards fine grained human behaviour simulation models. In: *2016 Winter Simulation Conference (WSC)*. pp. 3452–3463. IEEE (2016)
46. Song, Y.P., Wu, X., Yuan, Z., Qiao, J.J., Peng, Q.: Posturehmr: Posture transformation for 3d human mesh recovery. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9732–9741 (2024)
47. Stathopoulos, A., Han, L., Metaxas, D.: Score-guided diffusion for 3d human recovery. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 906–915 (2024)
48. Toivonen, M., Saari, E.: *Human-centered digitalization and services*, vol. 19. Springer (2019)
49. Tripathi, S., Müller, L., Huang, C.H.P., Omid, T., Black, M.J., Tzionas, D.: 3D human pose estimation via intuitive physics. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4713–4725 (2023), <https://ipman.is.tue.mpg.de>

50. wang, t., Zhang, J., Cai, Y., Yan, S., Feng, J.: Direct multi-view multi-person 3d pose estimation. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) *Advances in Neural Information Processing Systems*. vol. 34, pp. 13153–13164. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/6da9003b743b65f4c0ccd295cc484e57-Paper.pdf
51. Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., Wang, W., Zhou, S., Lin, G., Fu, Y., et al.: Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475* (2017)
52. Xu, L., Jin, S., Liu, W., Qian, C., Ouyang, W., Luo, P., Wang, X.: Zoomnas: Searching for whole-body human pose estimation in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022)
53. Yu, Z., Zhang, L., Xu, Y., Tang, C., Tran, L., Keskin, C., Park, H.S.: Multiview human body reconstruction from uncalibrated cameras. *Advances in Neural Information Processing Systems* **35**, 7879–7891 (2022)
54. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 11446–11456 (2021)
55. Zhang, Y., Wang, Z., Zhang, J., Shan, G., Tian, D.: A survey of immersive visualization: Focus on perception and interaction. *Visual Informatics* **7**(4), 22–35 (2023)
56. Zhang, Z., Wang, C., Qiu, W., Qin, W., Zeng, W.: Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild. *International Journal of Computer Vision* **129**(3), 703–718 (2021)
57. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 5745–5753 (2019)