Computers & Graphics (2025)

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Appearance as Reliable Evidence: Reconciling Appearance and Generative Priors for Monocular Motion Estimation

Zipei Chen[a,1], Yumeng Li[a,1], Zhong Ren[a], Yao-Xiang Ding[a,*], Kun Zhou[a]

[a] *State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China*

ARTICLE INFO

Article history:

human pose estimation, motion prior, appearance modeling, 3D Gaussian splatting

ABSTRACT

Monocular motion estimation in real scenes is challenging with the presence of noisy and possibly occluded detections. The recent method proposes to introduce a diffusion-based generative motion prior, which treats input detections as noisy partial evidence and generates motion through denoising. This advances robustness and motion quality, yet regardless of whether the denoised motion is close to visual observation, which often causes misalignment. In this work, we propose to reconcile model appearance and motion prior, which enables appearance to play the crucial role of providing reliable noise-free visual evidence for accurate visual alignment. The appearance is modeled by the radiance of both scene and human for joint differentiable rendering. To achieve this with monocular RGB input without mask and depth, we propose a semantic-perturbed mode estimation method to faithfully estimate static scene radiance from dynamic input with complex occlusion relationships, and a polyline depth calibration method to leverage knowledge from depth estimation model to recover the missing depth information. Meanwhile, to leverage knowledge from motion prior and reconcile it with the appearance guidance during optimization, we also propose an occlusion-aware gradient merging strategy. Experimental results demonstrate that our method achieves better-aligned tracking results while maintaining competitive motion quality. Code will be open-sourced upon acceptance.

1. Introduction

Motion estimation plays a crucial role in numerous real-world applications ranging from augmented reality to robotics [1, 2, 3]. Existing regression-based methods directly estimate motion from RGB input [4, 5, 6], while optimization-based methods fit motion to detected landmarks [7, 8, 9, 10]. However, in many scenarios, it is required to conduct motion estimation using monocular RGB-only or RGB-D input under significant detection noise and occlusion, which is a particularly challenging and heavily ill-posed situation, making the above methods difficult to handle [11]. Recently, an insightful approach RoHM [11] is proposed to utilize a diffusion-based generative model to effectively learn the human motion distribution from large-scale dataset. By performing denoising during the estimation process, this method is robust against noisy input and can complete the motion even with incomplete detections, thus producing motions with superior quality.

Despite its robustness, this approach still struggles with accurately aligning estimated motion with visual observation (Fig. 1). The cause is that it still relies on noisy detections as the sole evidence. Through denoising, the generative motion

*Corresponding author.
e-mail: 12421191@zju.edu.cn (Zipei Chen), yumeng.li@zju.edu.cn (Yumeng Li), renzhong@zju.edu.cn (Zhong Ren), dingyx.gm@gmail.com (Yao-Xiang Ding), kunzhou@acm.org (Kun Zhou)

[1] These authors contributed equally to this work.

Fig. 1. Comparison on the results of monocular motion estimation between our method and motion-prior-only approach RoHM [11]. By reconciling appearance and motion prior, Our method achieves better-aligned tracking results while maintaining competitive motion quality even under heavy occlusion.

prior finds a plausible motion in learned distribution near the noisy input in an unconditional manner, which is not necessarily more accurate.

To address this, we propose to reconcile model appearance and motion prior, leveraging human appearance as a reliable, noise-free evidence to align estimated motion with visual observations. For appearance, we model 3D radiance of the scene and perform differentiable rendering for scene and human radiance together. From rendered output, the input image can serve as a robust guidance for motion estimation. However, in the monocular setting, the human appearance is heavily entangled with the scene and often occluded, with no accurate mask and depth information available, making radiance estimation a challenging task.

To tackle these challenges, we propose two essential techniques for scene-human separation and depth estimation. For scene-human separation, we propose a semantic-perturbed mode estimation approach to faithfully estimate static scene radiance from dynamic input. Mode estimation enables to find the most-occured element for the value distribution for a pixel across time. To account for cases where the human occupies the location longer than the scene (e.g., a person standing at a location during most of the video), we leverage predicted noisy segmentation through a semantic perturbation mechanism. By adding noises to predicted human segmentation before mode estimation, the mode of human region in distribution is lowered and spread out in a soft manner, which accounts for the noisy nature of the predictions. Furthermore, to obtain more accurate depth information, we propose a polyline depth calibration approach to leverage knowledge from off-the-shelf depth estimation models to recover the missing depth information, which enables nonlinearly adjustment of the depth scale.

In addition, we propose an optimization technique to reconcile the information from appearance and motion prior through an occlusion-aware gradient merging strategy. When we can directly see a body part, we trust the appearance more, and selects appearance gradient as primary gradient. When a body part is occluded, we trust the motion prior more and selects motion prior gradient as primary gradient. With primary gradient determining the overall gradient direction, we subsequently perform a neighborhood search to determine the final gradient descendant direction to minimize the upper bound for both objectives, aligning the tracked motion with the input video while preserving motion quality.

Experimental results demonstrate that our approach makes estimated motion significantly more aligned with visual observation, while retaining comparable motion quality. We also find that existing metrics can be saturated due to recent improvements in motion quality and imperfect ground truth annotation. To rigorously assess the alignment between the tracked motion and video content, we also manually annotated human masks and propose to use the intersection over union (IoU) metric in evaluation to better reflect the alignment between estimated motion and human observation. Our code and annotated mask data for OMR calculation will be released upon acceptance.

2. Related Work

Monocular motion estimation is a long-standing problem with various approaches. We follow previous literature and discuss them in three categories: *Regression-based*, *Optimization-based*, and the recently proposed method leveraging generative models, which we classify into *Generative prior-based*.

Regression-based. Regression-based methods directly estimate human shape and body pose given monocular input. A line of work focuses on human mesh recovery (HMR), which regresses a template human mesh representation [12, 13] from a single image [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 16, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35], which can be naturally extended to motion by estimating video frame-by-frame. Recent HMR works also take robustness, e.g. occlusion, into consideration [36, 37, 38, 39, 40, 41, 42]. Despite being applicable, they do not model the temporal distribution of human motion. Another line of work directly regresses local human motion given a sequence of inputs [43, 44, 4, 5, 6, 45, 46, 47, 48, 49, 50, 51, 52, 53], where recent work further expands this to estimate global motion with cameras or world coordinate systems [2, 10, 54, 55, 56]. These methods are more effective for tracking motion in the wild with moving cameras,

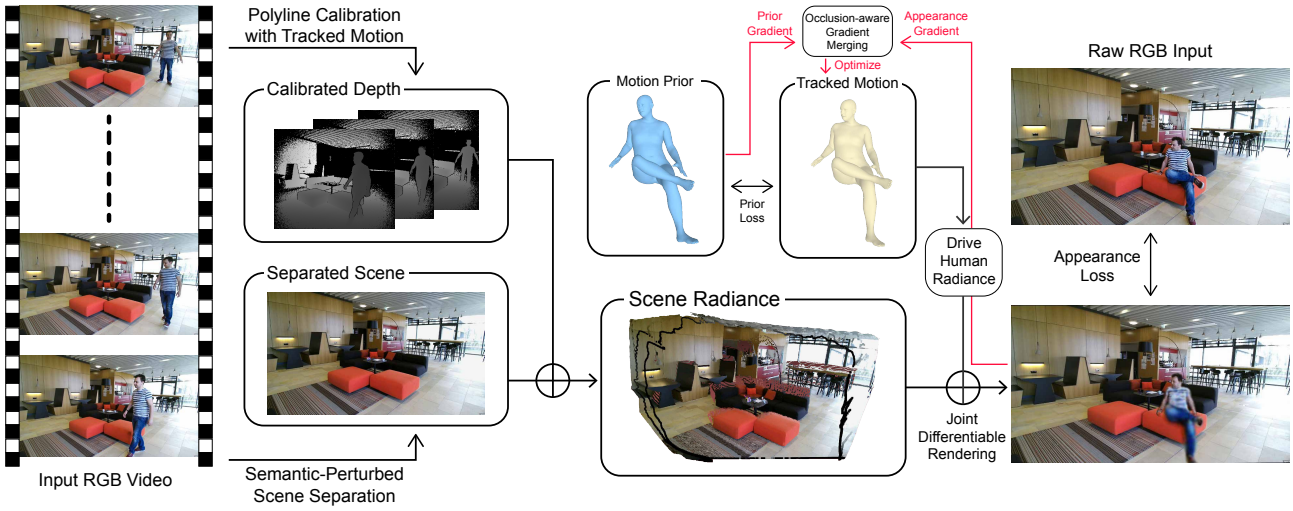

Fig. 2. The overall pipeline of our method. We estimated depth from RGB-only video and calibrate it with our motion, meanwhile performing semantic-perturbed scene separation. We construct a static scene radiance using monocular appearance and depth. The motion drives human radiance, which is jointly rendered with scene radiance, and directly compute loss with the raw input to optimize the predicted motion. During optimization, a motion prior provides consistent guidance. These two gradients are merged in an occlusion-aware manner to balance two different objectives, generating the final gradient. Under RGB-D setting, the calibrated depth is replaced with input depth.

while remaining less reliable when tracking heavy occlusion in monocular video based on indoor settings [11], which is what our method focuses on.

Optimization-based. Optimization-based methods perform motion tracking by minimizing the loss computed with detections, commonly including landmarks or silhouettes [57, 58, 59, 60, 7, 8, 9, 10], or for motion refinement using energy terms such as smoothness or foot-skating losses [61, 62, 57]. Some previous work incorporates appearance constraints to refine the motion [57, 58], which are paired with accurate silhouettes or accurate initial tracking results from inertial motion capture systems. Our method proposes to leverage appearance as visual evidence to rectify the generative prior in challenging real-world scenarios, which requires handling heavy occlusion without accurate masks while improving motion quality by reconciling with the generative motion prior.

Generative prior-based. With the recent advancements in powerful generative models in human motion via diffusion [63, 64, 65, 66, 67, 68, 69, 70], a notable recent work RoHM [11] explores applying diffusion to motion estimation in challenging real-world scenarios. By leveraging powerful diffusion models, RoHM shows promising performance against heavy occlusions and noisy inputs while modeling long-term temporal correspondence compared to previous motion priors [7, 8]. However, by assuming the input of the diffusion model as noisy evidence, the denoised motion is not guaranteed to be more aligned with the visual observation. Our work is an optimization-based method that is closely related to RoHM. More importantly, we further aim to achieve accurate alignment with challenging real-world input while benefiting from powerful generative models for superior motion quality.

3. Overview

As illustrated in Fig. 2, the pipeline of our approach takes monocular RGB(-D) (i.e. RGB-only or RGB-D) input from a static camera with known intrinsics and extrinsics, and estimates human motion in 3D space. The pipeline consists of two major parts: radiance modeling and motion estimation.

Radiance modeling. First, the *scene-human separation* is conducted. We separate the static scene from the dynamic input and obtain an initial mode estimation using the semantic-perturbed mode estimation method. Afterwards, the *depth calibration* is performed if depth information is not pre-given, which is done by matching the 3D human motion via a polyline adjustment. The scene radiance is constructed using the separated static scene and depth from an estimation model. Finally, the scene radiance is rendered differentiably alongside the human radiance.

Motion estimation. Motion estimation is obtained from a first-order optimization procedure. Through differentiable rendering, we optimize the motion estimation by computing the loss directly from the raw RGB input and back-propagating gradients to the human radiance that is bound to the template mesh, thereby refining the estimated motion. To reconcile with the diffusion motion prior, we constrain the joint locations and velocities from deviating excessively from the motion predicted by the prior, thereby generating an additional gradient. These two gradients are dynamically merged using an occlusion-aware strategy that relies more on the motion prior when a body region's appearance is less visible. Finally, a neighborhood search determines the final gradient descent direction to minimize the upper bound for both objectives.

4. Radiance Modeling

We model the radiance of both the scene and the human using 3D Gaussian Splatting [71]. This approach enables direct loss computation with the input RGB video without relying on an accurate mask, which is difficult to acquire under our challenging settings. The 3D Gaussians representing the scene remain static in 3D space, whereas the Gaussians modeling the human radiance are bound to the mesh of the estimated motion so that the motion can be corrected using the gradients back-propagated from the radiance. Constructing 3D Gaussian on the canonical space of the human mesh can also be seen in recent Gaussian-based human avatars [72, 73].

4.1. Scene Radiance

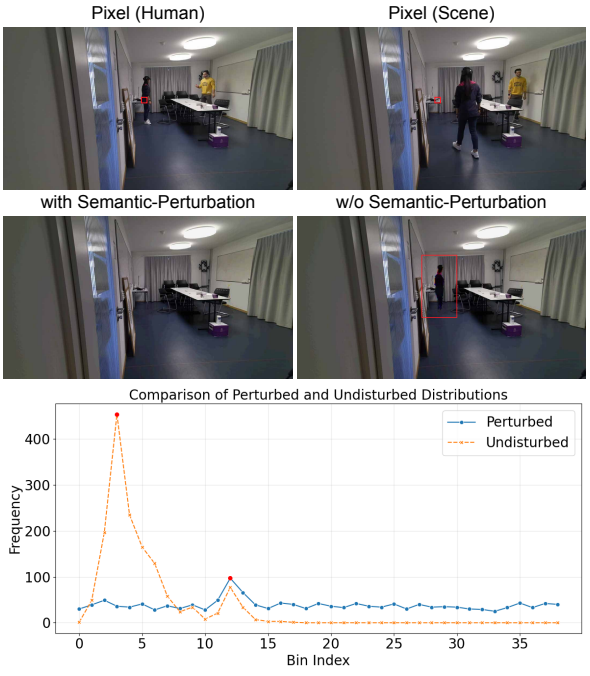

Fig. 3. Qualitative demonstration of semantic perturbation. We address cases where a pixel is occupied by humans for a longer duration than by the static background. By using an estimated mask to apply noise to the input, we spread out the frequency of human presence in the distribution of one pixel (highlighted with enlarged red square for visualization) and allow the static scene to stand out. Distribution is handled in a discrete manner by grouping into bins.

Given a monocular RGB(-D) video and an initial motion estimation, we separate the static 3D scene from the dynamic input, estimate the depth, and calibrate the estimated depth to match the motion so as to correctly model their relative occlusion relationships. The scene radiance is then constructed using both the depth and the RGB appearance. When the input is given in the RGB-D setting, the depth is used directly for both the initial motion estimation and the scene radiance modeling.

Scene-human separation. To separate the static scene from the dynamic input, we exploit the statistical characteristics of the static background, which remains largely consistent throughout the video, and use mode estimation to isolate it. Let $I_t(x, y)$ denote the pixel intensity at location (x, y) in frame t, for $t = 1, \ldots, T$, where T is the total number of frames in the sequence. We define the static scene signal $S(x, y)$ at each pixel as:

$$S(x, y) = \text{mode}\{I_t(x, y) \mid t = 1, \ldots, T\}. \tag{1}$$

this mode selection enables us to extract the static scene from the video.

However, if a human remains static for an extended period, the human region may erroneously dominate the distribution. In such cases, the scene radiance might capture the appearance of the human, preventing the image loss from effectively back-propagating gradients to the human radiance and thereby hindering motion correction. To mitigate this issue, we leverage noisy semantic segmentation as prior guidance. Specifically, we use an off-the-shelf instance segmentation method [74] to detect human regions, and then perturb these regions with random noise. This perturbation dilutes the mode in regions consistently identified as human, allowing the true static background to emerge as the dominant mode in the overall distribution.

Scene modeling with depth calibration. Given monocular RGB inputs from a static camera, if the depth information is missing, we lack direct 3D structure information and cannot employ methods like structure-from-motion (SfM) [75]. Therefore, to construct a 3D scene radiance that interacts correctly with the 3D motion, we first estimate the depth using an off-the-shelf depth estimation model [74]. Although this estimated depth provides reliable relative depth relationships (i.e., which objects are closer or farther), it is generally misaligned with the 3D motion when used directly in the same coordinate system.

To align the estimated depth with the 3D motion, we leverage the relationship between the full depth map, denoted by D, and the depth of the motion region, denoted by D^m. Here, D^m is defined only for the subset of pixels Ω^m that belong to the motion region. Because the relationship between these depths is non-linear, we propose to calibrate the estimated depth using a polyline transformation.

Specifically, we first cluster the estimated depth map D into K clusters and initialize K learnable parameters corresponding to the cluster centers, denoted as $\{d_k\}_{k=1}^K$. For a given pixel $i \in \Omega^m$, its calibrated depth is represented by a polyline interpolation:

$$D_i = d_k + (d_{k+1} - d_k) \cdot w_i, \tag{2}$$

where d_k and d_{k+1} are the adjacent cluster center values that pixel i falls between, and $w_i \in [0, 1]$ is the interpolation weight.

We then optimize these parameters by minimizing the following loss function over the motion region Ω^m:

$$L_{\text{depth}} = \sum_{i \in \Omega^m} \left| D_i - D_i^m \right|, \tag{3}$$

where D_i is the calibrated depth value for pixel i from the full depth map, and D_i^m is the corresponding depth obtained from the motion estimation. This formulation ensures that only the depth values in the motion region are aligned, preserving the relative depth relationships while correcting the misalignment between the full depth and the motion depth. We note that our

method shares a similar spirit with classic lookup table (LUT) methods used in color correction [76].

With known camera poses, intrinsics, and the calibrated depth, we can estimate the 3D radiance. We initialize 3D Gaussians in space using the calibrated depth and optimize them to fit the separated static scene, thereby obtaining the 3D scene radiance. In scenarios where the depth is provided, we directly use the given depth.

4.2. Human Radiance

We model the human radiance using a set of 3D Gaussians anchored in the canonical space of the SMPL model [12]. To focus on optimizing motion parameters rather than free-floating radiance, we restrict the degrees of freedom of these Gaussians. Specifically, each Gaussian is parameterized by a scaling factor $s \in \mathbb{R}^3$, a rotation represented by a quaternion $q \in \mathbb{R}^4$, and color $f \in \mathbb{R}^3$.

Note that the Gaussians are tied exclusively to the human mesh. That is, the Gaussians in world space in frame t is given by:

$$\mathbb{G}_x^h(t) = M\left(\gamma_t, \Phi_t, \theta_t, \beta\right), \tag{4}$$

where $M(\cdot)$ denotes the linear blend skinning which transforms motion to 3D mesh vertices. γ_t and Φ_t represent the global translation and global orientation at time t, respectively, θ_t encodes the body pose, and β denotes the body shape parameters.

By binding the Gaussian to the mesh, any adjustments in radiance location through the optimization of the pixel-wise RGB appearance loss L_{RGB} directly refine the underlying motion. In other words, the Gaussians cannot "float" independently; their positions are solely updated by refining the motion parameters so that appearance alignment drives motion estimation.

5. Motion Estimation

We now perform motion estimation guided by both appearance and a motion prior. With the scene radiance and an initial motion estimation available, we initialize the human radiance on the human mesh, differentiable render the scene and human radiance, and compute a loss with the raw RGB input. Back-propagating the gradients allows us to optimize both the human radiance and its underlying motion. To maintain motion quality, we incorporate guidance from a motion prior via a gradient merging strategy. In the following sections, we introduce the appearance-based (radiance) optimization objective, the motion prior guidance objective, and then explain how these are reconciled into the final optimization objective.

5.1. Radiance Optimization Objective

Given an RGB video sequence $\{I_t\}_{t=1}^T$, the scene radiance $\mathcal{G}^s$, the initial human motion parameters $\mathcal{M}$, and the human radiance represented by a set of 3D Gaussians $\mathcal{G}^h$ in the canonical space, we first transform the human Gaussians to the world space using the mapping in Equation (5):

$$\mathbb{G}_x^h(t) = M\left(\gamma_t, \Phi_t, \theta_t, \beta\right), \tag{5}$$

where $M(\cdot)$ is the SMPL skinning transformation, and γ_t, Φ_t, θ_t, and β denote the global translation, global orientation, body pose, and body shape, respectively.

We jointly render $\mathcal{G}^s$ and $\mathcal{G}^h$ through differentiable Gaussian rasterization to obtain the rendered frame $\hat{I}_t$. The appearance loss is then defined as

$$L_{\text{RGB}} = \sum_{t=1}^{T} \|\hat{I}_t - I_t\|_2^2. \tag{6}$$

This loss yields gradients on both the Gaussian appearance properties (e.g., color, scaling, rotation) and the motion parameters $\mathcal{M}$. The gradient with respect to the motion parameters is later reconciled with that from the motion prior guidance.

5.2. Prior Guidance Objective

Appearance-based optimization alone may yield implausible motion, especially under occlusion, because it does not account for the motion distribution. To improve motion quality, we incorporate a motion prior that provides reference joint velocities and positions. We define the reference loss as a combination of a velocity loss L_v and a joint position loss L_j:

$$L_{\text{ref}} = L_v + L_j = \|v_{\text{prior}} - v_p\|_1 + \|j_{\text{prior}} - j_p\|_1, \tag{7}$$

where:

- v_{prior} and j_{prior} are the joint velocities and positions predicted by the motion prior,
- v_p and j_p are the corresponding velocities and positions computed from the current motion parameters $\mathcal{M}$.

In our implementation, we denoise the initial motion once to obtain the reference values v_{prior} and j_{prior} rather than re-sampling the reference at every optimization step.

Back-propagating L_{ref} produces a gradient on the motion parameters from the prior, denoted as $\nabla\mathcal{M}_{\text{prior}}$.

5.3. Final Optimization on Motion

To integrate the appearance and prior guidance gradients, we employ a gradient merging strategy that adapts based on occlusion. When a body part is visible, the appearance gradient is more reliable, while under occlusion, the prior gradient takes precedence.

Let $\Omega_j \subset \{1, \ldots, T\}$ denote the set of frames in which human joint j is visible, as determined by an occlusion mask. For joint j, let $\nabla\mathcal{M}_{\text{app}}^j$ denote the appearance gradient and $\nabla\mathcal{M}_{\text{prior}}^j$ the gradient from the motion prior. We define the dominant gradient for joint j as

$$\nabla\mathcal{M}_{\text{dom}}^j = \begin{cases} \nabla\mathcal{M}_{\text{app}}^j, & \text{if } t \in \Omega_j \text{ (joint } j \text{ is visible);} \\ \nabla\mathcal{M}_{\text{prior}}^j, & \text{otherwise.} \end{cases} \tag{8}$$

Subsequently, motivated by [77], we refine the final gradient by searching for an update direction in a neighborhood around the dominant gradient that minimizes the upper bound of the

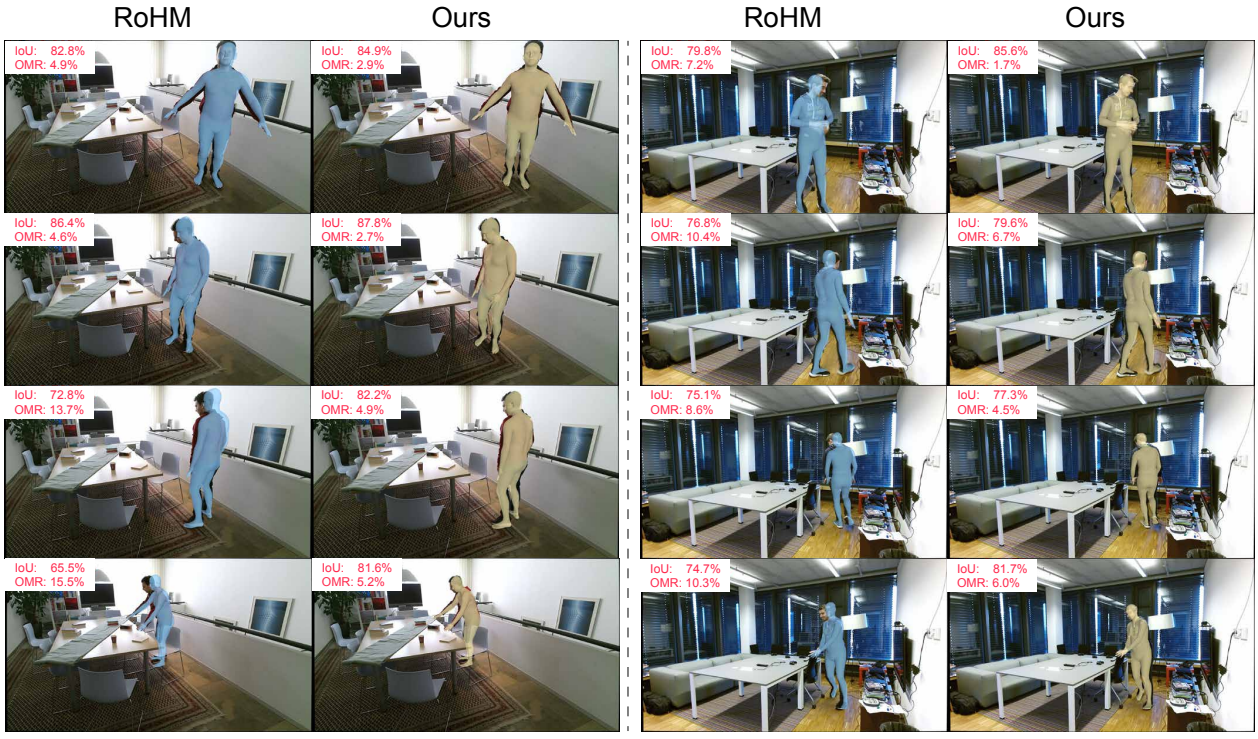

Fig. 4. Qualitative comparison with RoHM. Corresponding Intersection-over-Union (IoU) and Out-of-Mask Rate (OMR) are marked on each frame. The results include cases where we made less improvements. We demonstrate that OMR is a robust metric since the estimated mesh should not fall outside the ground truth mask, and IoU also aligns with human observation. As shown in the images, a decrease of around 4% in OMR in one frame can result in a visually noticeable improvement. More results are presented in the supplementary material.

two objectives. Formally, for each joint j, the final gradient is given by

$$\nabla\mathcal{M}^j = \underset{\nabla\in\mathcal{N}\left(\nabla\mathcal{M}^j_{\text{dom}}\right)}{\operatorname{argmin}} \max\left\{L_{\text{RGB}}(\mathcal{M}_{-\nabla}),\ L_{\text{ref}}(\mathcal{M}_{-\nabla})\right\}, \quad (9)$$

where $\mathcal{N}\left(\nabla\mathcal{M}^j_{\text{dom}}\right)$ denotes a neighborhood of the dominant gradient direction for joint j, $\mathcal{M}_{-\nabla}$ denotes the parameter updated by ∇. The final gradient $\nabla\mathcal{M}^j$ is then used by the optimizer to update the motion parameters in one optimization step. The iterative optimization process follows this strategy until convergence.

6. Experiments

We try to answer the following questions from the experiments: 1) Whether our method can achieve better tracking results meanwhile maintain comparable performance on motion quality; 2) Whether our proposed metrics can reflect accuracy of motion estimation when given metrics are saturated; 3) Whether the sub-modules of our proposed pipeline are effective and necessary.

6.1. Datasets, Baselines, and Metrics

Datasets and baselines. Following previous work [11], we conduct experiments on two benchmark datasets to verify the effectiveness of our method. The PROX dataset [78] consists of multiple single-person monocular RGB(-D) videos capturing 3D human–scene interactions across a range of indoor environments. Similarly, the EgoBody dataset [79] comprises several multi-person motion RGB sequences recorded in indoor environments. In both datasets, the cameras remain fixed throughout each video sequence. We evaluate our method on a subset of the occluded qualitative data, as in previous work [7, 11]. Note that there is no ground truth motion in PROX. The EgoBody dataset [79] provides indoor video sequences with ground truth SMPL parameter annotations.

Since various motion estimation methods prioritize different settings, we largely follow our closest work, RoHM [11], in our baseline choices for each dataset, with details listed in Tab. 2, 3, and 1. VPoser-t represents the intermediate results from HuMor [7]. For our ablation study, as described in Tab. 4, we experiment with removing: (1) scene radiance (SR), which uses the estimated mask instead of constructing scene radiance; (2) depth calibration (DC), which does not calibrate the depth to align with motion in the world coordinate system; (3) the diffusion prior (DP), which removes diffusion prior guidance during optimization; and (4) gradient merging (GM), which disregards occlusion and directly merges the loss terms with weights adjusted to maximize IoU and OMR while preserving competitive motion quality.

Metrics. We first introduce the established metrics used in previous works. Accuracy is measured by computing the Mean Per Joint Position Error in both the pelvis ($MPJPE$) and global

Table 1. Quantitative comparison results on our newly proposed metrics OMR and IoU in percentage. LEMO is only designed for input with ground truth depth. Our method shows consistent improvement. See Fig. 4 for qualitative results.

Method	PROX (RGB-Only)		PROX (RGB-D)		EgoBody	
	IoU↑	OMR↓	IoU↑	OMR↓	IoU↑	OMR↓
VPoser-t	66.74	24.95	72.90	21.0	71.75	14.6
LEMO	-	-	75.91	7.98	-	-
RoHM	73.11	10.71	75.76	9.03	72.12	14.80
Ours	**78.26**	**6.90**	**79.88**	**5.81**	**75.06**	**12.11**

Table 2. Quantitative comparison results on motion quality metrics on PROX, including skating ratio, acceleration (m/s^2), foot penetration distance (mm). There is no ground truth motion, so comparing GMPJPE and MPJPE is impossible.

Method	PROX (RGB-Only)			PROX (RGB-D)		
	Skat↓	Acc↓	Dist↓	Skat↓	Acc↓	Dist↓
CLIFT [24]	0.70	49.60	61.80	-	-	-
VPoser-t	0.21	3.20	50.14	0.28	3.40	48.75
HuMor [7]	0.13	2.30	35.41	0.11	1.9	54.76
LEMO [9]	-	-	-	0.17	**1.8**	34.22
PhaseMP [8]	0.18	**1.8**	46.96	-	-	-
RoHM [11]	**0.11**	2.20	**9.77**	**0.04**	1.9	**3.36**
Ours	0.13	2.84	13.22	0.08	2.35	7.14

coordinate systems (*GMPJPE*); both metrics are expressed in millimeters. The plausibility of the estimated motion is measured using metrics such as foot skating ratio (*Skat*), mean per joint acceleration (*acc*, measured in m^2/s), and the mean toe joint penetration distance (*Dist*, in millimeters). Following previous work [11], foot skating is defined under the condition where the distance between the foot joints and the ground is less than 10 cm while the foot joints exhibit velocities exceeding 10 cm/s.

These metrics are effective when motion quality and accuracy are low; however, with recent advancements, the metrics have become saturated and struggle to reflect the actual perceived motion quality when the results are sufficiently good. The underlying reasons are straightforward: 1) Error with Ground Truth (*MPJPE*, *GMPJPE*): Monocular motion estimation of clothed humans is ill-posed since the ground-truths are annotated for naked humans; furthermore, the ground truth motion is manually annotated, so it only guarantees to be accurate when the precision requirement is not high; For these reasons, when the estimated motion is close enough to the ground truth, this metric is saturated; 2) Acceleration (*Acc*): Acceleration can clearly indicate unnatural twitching when it is very high, but when it is already low, the metric saturates because lower is not necessarily better, since the best result would not be static motion; 3) Foot metrics (*Skat*, *Dist*): The skating is calculated based on a preset rule, where a human might simply not raise the feet high enough when moving, and the ground penetration difference is only a few millimeters.

To better reflect the accuracy of the estimated motion in alignment with human perception, we propose a new supplementary metric called Out of Mask Ratio (*OMR*), paired with the classic Intersection over Units (*IoU*) metric. The motivation is that the predicted unclothed human motion should strictly remain within the human region; any deviation outside the human mask can be considered erroneous. The *IoU* metric calculates the degree to which the human fills the mask, thereby reflecting the alignment between the estimated motion and the clothed human to a fair extent. We manually annotate an accurate ground truth mask, and the *OMR* is calculated as the percentage of the estimated human motion that falls outside the ground truth mask relative to the size of the mask. The *IoU* metric measures the intersection between the human motion and the ground truth mask. To handle occlusion, we specifically annotate an occlusion mask whenever occlusion occurs, and human motion covered by the occlusion mask is excluded from the calculation to ensure fairness. As shown in our qualitative results in Fig. 4, we find that the newly proposed metric is intuitive and aligns well with human perception; approximately a 4% difference in *OMR* in one frame can yield a visually noticeable improvement.

Table 3. Quantitative comparison results on established metrics on Egobody. Our method achieves competitive quantitative results.

Method	EgoBody				
	Skat↓	Acc↓	Dist↓	GMPJPE↓	MPJPE↓
VPoser-t	0.15	3.63	11.62	287.30	81.39
HuMor [7]	0.14	3.50	17.44	340.3	92.63
RoHM [11]	**0.02**	**2.63**	**2.98**	283.04	74.71
Ours	0.03	**2.63**	3.31	**281.11**	**73.50**

6.2. Results and Discussions

Implementation details. Our proposed method is implemented in PyTorch on single NVIDIA 4090 GPU. We utilize the off-the-shelf diffusion motion prior from RoHM [11]. The motion estimation process with radiance and motion optimization is conducted for 100K steps per sequence, starting with an initial learning rate of 2×10^{-4}. The learning rate is halved every 20K steps, using a higher rate initially to explore the solution space and then decreasing it to converge on a better result. The motion is initialized using RoHM. We will open-source our code upon acceptance.

Motion quality. As described in Sec. 6.1, we evaluate motion quality using established metrics. As demonstrated in Tab. 2 and Tab. 3, our method achieves competitive motion quality, with comparable foot skating, acceleration, and ground penetration differences in a few millimeters.

Motion accuracy. As shown in Tab. 3, we compute GMPJPE and MPJPE on the EgoBody dataset, which provides ground truth motion. Our results are comparable to those of RoHM, with differences in a few millimeters. As discussed in Sec. 6.1, we argue that, similar to the motion quality metrics, GMPJPE and MPJPE become saturated. Therefore, we also present qualitative and quantitative results using OMR and IoU (see Fig. 4 and Tab. 1). We demonstrate that our improvements in visual alignment are significant. When GMPJPE and MPJPE are saturated, OMR and IoU serve as reliable metrics to reflect improvements in alignment accuracy. These metrics are also applicable to datasets without ground truth motion, such as PROX.

Fig. 5. Without dynamically merging gradients based on occlusion of joints, if we maintain a high gradient from the motion prior to preserve motion quality, it will also prevent the appearance module from correcting the motion when the gradients from the two sources conflict.

Table 4. Ablation study of our method on PROX, see details in Sec. 6.1. Using scene radiance instead of estimated mask yields significant increase in OMR. Compared with ablation baselines without motion prior or without occlusion-aware gradient merging, our full methods balances motion quality and accuracy.

Method	PROX				
	Skat↓	Acc↓	Dist↓	IoU↑	OMR↓
w/o SR	0.15	4.17	14.91	77.71	8.98
w/o DC	0.13	3.07	11.36	74.41	9.28
w/o DP	0.42	10.22	19.95	**80.84**	**4.37**
w/o GM	**0.09**	**2.79**	**10.46**	76.53	8.45
Ours	0.13	2.84	13.22	78.26	6.90

As shown in Fig. 4, an improvement of roughly 4% in OMR for one frame results in notably better alignment. Considering that there are many easy-to-estimate frames in the video, after averaging throughout the sequence, our quantitative results in Tab. 1 consistently show improved alignment over the baselines.

Ablations. As shown in Tab. 4, our full method strikes a balance between motion quality and accuracy. Using masks predicted from segmentation models instead of jointly modeling the scene and human radiance leads to inferior results, as segmentation is unreliable under heavy occlusion. Without depth calibration, the scene radiance is ineffective and results in the worst OMR. When removing the diffusion prior (DP) and removing gradient merging (GM), our method maintains competitive motion quality while achieving better accuracy, coming close to the results obtained using appearance guidance alone. Note that in the ablation experiment without occlusion-aware gradient merging, we tuned the loss weight to preserve competitive motion quality and see how the motion accuracy is. Without occlusion-aware merging to dynamically adjust the primary gradient, when gradients from appearance and the prior conflict, the appearance module struggles to fully correct the motion. Qualitative results are shown in Fig. 5.

Fig. 6. Failure Cases: Our method is not able to correct cases where the initially estimated motion is entirely incorrect, such as when the front and back are flipped.

Table 5. The comparison results of mean time cost (s) per frame between the SOTA regression-based, optimization-based method and our method (where *Reg*, *Opt* represent the Regression-based and Optimization-based method).

Method	Time Cost (s)
RoHM(*Reg*) [11]	**0.14**
HuMor(*Opt*) [7]	6.2
Ours	5.4

7. Limitations and Future Work

As shown in Fig. 6, our method cannot handle cases where the initial motion is entirely wrong. Besides, as you can see in Tab. 5, although 3D Gaussian is an effective radiance representation, similar to previous optimization-based works, our method is more time-consuming compared to regression methods. Moreover, owing to the assumption inherent in our scene radiance modeling, our method is confined to the scenarios with static cameras. When applied to footage captured by a moving camera, the radiance may model incorrectly with only monocular video, which leads to faulty occlusion reasoning and consequently produces misaligned results, as illustrated in Fig 7. Future work could explore feed-forward approaches that directly estimate human motion while being more aware of occlusions and human motion distribution.

8. Conclusion

In this paper, by taking appearance as a reliable visual evidence, we propose a monocular motion estimation method which reconcile model appearance and motion prior. To achieve this with only monocular video, a semantic-perturbed mode estimation and a polyline depth calibration method is proposed. Moreover, to leverage knowledge from motion prior and reconcile it with the appearance guidance in occluded scenarios, we also design an occlusion-aware gradient merging strategy. The extensive experiment results show that our method achieves more accurate visual alignment while preserves motion quality.

References

[1] T. He, J. Gao, W. Xiao, Y. Zhang, Z. Wang, J. Wang, Z. Luo, G. He, N. Sobanbabu, C. Pan, Z. Yi, G. Qu, K. Kitani, J. Hodgins, L. J. Fan, Y. Zhu, C. Liu, G. Shi, Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills, arXiv preprint arXiv:2502.01143 (2025).

[2] S. Shin, J. Kim, E. Halilaj, M. J. Black, WHAM: Reconstructing world-grounded humans with accurate 3D motion, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2024.

Fig. 7. Failure cases in scenarios of moving camera.

[3] M. Kocabas, Y. Yuan, P. Molchanov, Y. Guo, M. J. Black, O. Hilliges, J. Kautz, U. Iqbal, Pace: Human and camera motion estimation from in-the-wild videos, in: 2024 International Conference on 3D Vision (3DV), IEEE, 2024, pp. 397–408.

[4] M. Kocabas, N. Athanasiou, M. J. Black, Vibe: Video inference for human body pose and shape estimation, in: CVPR, 2020.

[5] H. Choi, G. Moon, J. Y. Chang, K. M. Lee, Beyond static features for temporally consistent 3d human pose and shape from a video, in: CVPR, 2021.

[6] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, J. Malik, Humans in 4D: Reconstructing and tracking humans with transformers, in: ICCV, 2023.

[7] D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, L. J. Guibas, HuMoR: 3D human motion model for robust pose estimation, in: ICCV, 2021.

[8] M. Shi, S. Starke, Y. Ye, T. Komura, J. Won, PhaseMP: Robust 3D pose estimation via phase-conditioned human motion prior, in: ICCV, 2023.

[9] S. Zhang, Y. Zhang, F. Bogo, M. Pollefeys, S. Tang, Learning motion priors for 4d human body capture in 3d scenes, in: ICCV, 2021.

[10] V. Ye, G. Pavlakos, J. Malik, A. Kanazawa, Decoupling human and camera motion from videos in the wild, in: CVPR, 2023.

[11] S. Zhang, B. L. Bhatnagar, Y. Xu, A. Winkler, P. Kadlecek, S. Tang, F. Bogo, Rohm: Robust human motion reconstruction via diffusion, in: CVPR, 2024.

[12] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M. J. Black, SMPL: A skinned multi-person linear model, ACM Trans. Gr. 34 (6) (2015) 1–16.

[13] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, M. J. Black, Expressive body capture: 3D hands, face, and body from a single image, in: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10975–10985.

[14] S. K. Dwivedi, Y. Sun, P. Patel, Y. Feng, M. J. Black, TokenHMR: Advancing human mesh recovery with a tokenized pose representation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[15] A. Kanazawa, M. J. Black, D. W. Jacobs, J. Malik, End-to-end recovery of human shape and pose, in: CVPR, 2018.

[16] N. Kolotouros, G. Pavlakos, K. Daniilidis, Convolutional mesh regression for single-image human shape reconstruction, in: CVPR, 2019.

[17] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, P. V. Gehler, Unite the people: Closing the loop between 3D and 2D human representations, in: CVPR, 2017.

[18] Y. Xu, S.-C. Zhu, T. Tung, Denserac: Joint 3D pose and shape estimation by dense render-and-compare, in: ICCV, 2019.

[19] J. Zhang, D. Yu, J. H. Liew, X. Nie, J. Feng, Body meshes as points, in: CVPR, 2021.

[20] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, C. Lu, HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation, in: CVPR, 2021.

[21] K. Lin, L. Wang, Z. Liu, End-to-end human pose and mesh reconstruction with transformers, in: CVPR, 2021.

[22] M. Kocabas, C.-H. P. Huang, J. Tesch, L. Müller, O. Hilliges, M. J. Black, SPEC: Seeing people in the wild with an estimated camera, in: ICCV, 2021.

[23] J. Cho, K. Youwang, T.-H. Oh, Cross-attention of disentangled modalities for 3D human mesh recovery with transformers, in: ECCV, 2022.

[24] Z. Li, J. Liu, Z. Zhang, S. Xu, Y. Yan, Cliff: Carrying location information in full frames into human pose and shape estimation, in: ECCV, 2022.

[25] Q. Fang, Q. Shuai, J. Dong, H. Bao, X. Zhou, Reconstructing 3D human pose by watching humans in the mirror, in: CVPR, 2021.

[26] N. Kolotouros, G. Pavlakos, M. J. Black, K. Daniilidis, Learning to reconstruct 3D human pose and shape via model-fitting in the loop, in: ICCV, 2019.

[27] J. Song, X. Chen, O. Hilliges, Human body model fitting by learned gradient descent, in: ECCV, 2020.

[28] D. Wang, S. Zhang, 3d human mesh recovery with sequentially global rotation estimation, in: ICCV, 2023.

[29] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, M. J. Black, Collaborative regression of expressive bodies using moderation, in: 2021 International Conference on 3D Vision (3DV), 2021.

[30] J. Lin, A. Zeng, H. Wang, L. Zhang, Y. Li, One-stage 3d whole-body mesh recovery with component aware transformer, in: CVPR, 2023.

[31] K. Lin, L. Wang, Z. Liu, End-to-end human pose and mesh reconstruction with transformers, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 1954–1963.

[32] J. Cho, K. Youwang, T.-H. Oh, Cross-attention of disentangled modalities for 3d human mesh recovery with transformers, in: European Conference on Computer Vision, Springer, 2022, pp. 342–359.

[33] Z. Dou, Q. Wu, C. Lin, Z. Cao, Q. Wu, W. Wan, T. Komura, W. Wang, Tore: Token reduction for efficient human mesh recovery with transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15143–15155.

[34] J. Li, S. Bian, Q. Liu, J. Tang, F. Wang, C. Lu, Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12933–12942.

[35] H. Choi, G. Moon, K. M. Lee, Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16, Springer, 2020, pp. 769–787.

[36] Y. Zhang, P. Ji, A. Wang, J. Mei, A. Kortylewski, A. Yuille, 3d-aware neural body fitting for occlusion robust 3d human pose estimation, in: ICCV, 2023.

[37] J. Li, Z. Yang, X. Wang, J. Ma, C. Zhou, Y. Yang, Jotr: 3d joint contrastive learning with transformers for occluded human mesh recovery, in: ICCV, 2023.

[38] R. Khirodkar, S. Tripathi, K. Kitani, Occluded human mesh recovery, in: CVPR, 2022.

[39] Q. Liu, Y. Zhang, S. Bai, A. Yuille, Explicit occlusion reasoning for multi-person 3d human pose estimation, in: ECCV, 2022.

[40] M. Kocabas, C.-H. P. Huang, O. Hilliges, M. J. Black, PARE: Part attention regressor for 3D human body estimation, in: ICCV, 2021.

[41] S. Zhang, Q. Ma, Y. Zhang, S. Aliakbarian, D. Cosker, S. Tang, Probabilistic human mesh recovery in 3d scenes from egocentric views, in: ICCV, 2023.

[42] C. Rockwell, D. Fouhey, Full-body awareness from partial observations, in: ECCV, 2020.

[43] Y. Cheng, B. Yang, B. Wang, Y. Wending, R. Tan, Occlusion-Aware Networks for 3D Human Pose Estimation in Video, in: ICCV, 2019.

[44] A. Kanazawa, J. Y. Zhang, P. Felsen, J. Malik, Learning 3d human dynamics from video, in: CVPR, 2019.

[45] Y. Sun, Y. Ye, W. Liu, W. Gao, Y. Fu, T. Mei, Human mesh recovery from monocular images via a skeleton-disentangled representation, in: ICCV, 2019.

[46] Z. Luo, S. A. Golestaneh, K. M. Kitani, 3d human motion estimation via motion compression and refinement, in: ACCV, 2020.

[47] A. Zanfir, E. G. Bazavan, H. Xu, W. T. Freeman, R. Sukthankar, C. Sminchisescu, Weakly supervised 3d human pose and shape reconstruction with normalizing flows, in: ECCV, 2020.

[48] Y. You, H. Liu, T. Wang, W. Li, R. Ding, X. Li, Co-evolution of pose and mesh for 3d human body estimation from video, in: ICCV, 2023.

[49] H. Nam, D. S. Jung, Y. Oh, K. M. Lee, Cyclic test-time adaptation on monocular video for 3d human mesh reconstruction, in: ICCV, 2023.

[50] L. G. Foo, J. Gong, H. Rahmani, J. Liu, Distribution-aligned diffusion for human mesh recovery, in: ICCV, 2023.

[51] W.-L. Wei, J.-C. Lin, T.-L. Liu, H.-Y. M. Liao, Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video, in: CVPR, 2022.

[52] J. Rajasegaran, G. Pavlakos, A. Kanazawa, J. Malik, Tracking people by predicting 3d appearance, location and pose, in: CVPR, 2022.

[53] G. Pavlakos, J. Malik, A. Kanazawa, Human mesh recovery from multiple shots, in: CVPR, 2022.

[54] W. Yin, Z. Cai, R. Wang, F. Wang, C. Wei, H. Mei, W. Xiao, Z. Yang, Q. Sun, A. Yamashita, et al., Whac: World-grounded humans and cameras, in: European Conference on Computer Vision, Springer, 2024, pp. 20–37.

[55] Z. Shen, H. Pi, Y. Xia, Z. Cen, S. Peng, Z. Hu, H. Bao, R. Hu, X. Zhou, World-grounded human motion recovery via gravity-view coordinates, in: SIGGRAPH Asia 2024 Conference Papers, 2024, pp. 1–11.

[56] Y. Wang, Z. Wang, L. Liu, K. Daniilidis, Tram: Global trajectory and motion of 3d humans from in-the-wild videos, in: European Conference on Computer Vision, Springer, 2024, pp. 467–487.

[57] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, C. Theobalt, Livecap: Real-time human performance capture from monocular video, ACM Transactions On Graphics (TOG) 38 (2) (2019) 1–17.

[58] M. Kaufmann, J. Song, C. Guo, K. Shen, T. Jiang, C. Tang, J. J. Zárate, O. Hilliges, EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild, in: International Conference on Computer Vision (ICCV), 2023.

[59] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, M. J. Black, Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image, in: ECCV, 2016.

[60] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, M. J. Black, Expressive body capture: 3D hands, face, and body from a single image, in: CVPR, 2019.

[61] A. Arnab, C. Doersch, A. Zisserman, Exploiting temporal context for 3d human pose estimation in the wild, in: CVPR, 2019.

[62] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, C. Theobalt, Vnect: Real-time 3d human pose estimation with a single rgb camera, ACM Trans. Gr. 36 (4) (2017) 1–14.

[63] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, A. H. Bermano, Human motion diffusion model, in: ICLR, 2023.

[64] Y. Yuan, J. Song, U. Iqbal, A. Vahdat, J. Kautz, Physdiff: Physics-guided human motion diffusion model, in: ICCV, 2023.

[65] Y. Wang, Z. Leng, F. W. Li, S.-C. Wu, X. Liang, Fg-t2m: Fine-grained text-driven human motion generation via diffusion model, in: ICCV, 2023.

[66] M. Zhang, X. Guo, L. Pan, Z. Cai, F. Hong, H. Li, L. Yang, Z. Liu, Remodiffuse: Retrieval-augmented motion diffusion model, arXiv preprint arXiv:2304.01116 (2023).

[67] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, G. Yu, Executing your commands via motion diffusion in latent space, in: CVPR, 2023.

[68] K. Karunratanakul, K. Preechakul, S. Suwajanakorn, S. Tang, Guided motion diffusion for controllable human motion synthesis, in: ICCV, 2023.

[69] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, T. Chen, Motiongpt: Human motion as a foreign language, Advances in Neural Information Processing Systems 36 (2024).

[70] K. Karunratanakul, K. Preechakul, E. Aksan, T. Beeler, S. Suwajanakorn, S. Tang, Optimizing diffusion noise can serve as universal motion priors, in: CVPR, 2024.

[71] B. Kerbl, G. Kopanas, T. Leimkühler, G. Drettakis, 3d gaussian splatting for real-time radiance field rendering, ACM Transactions on Graphics 42 (4) (July 2023).

[72] Z. Qian, S. Wang, M. Mihajlovic, A. Geiger, S. Tang, 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting (2024).

[73] L. Hu, H. Zhang, Y. Zhang, B. Zhou, B. Liu, S. Zhang, L. Nie, Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[74] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587 (2017).

[75] S. Ullman, The interpretation of structure from motion, Proceedings of the Royal Society of London. Series B. Biological Sciences 203 (1153) (1979) 405–426.

[76] J. Selan, Using lookup tables to accelerate color transformations, in: GPU Gems 2, Addison-Wesley, 2005, pp. 381–392.

[77] B. Liu, X. Liu, X. Jin, P. Stone, Q. Liu, Conflict-averse gradient descent for multi-task learning, Advances in Neural Information Processing Systems 34 (2021) 18878–18890.

[78] M. Hassan, V. Choutas, D. Tzionas, M. J. Black, Resolving 3D human pose ambiguities with 3D scene constraints, in: ICCV, 2019.

[79] S. Zhang, Q. Ma, Y. Zhang, Z. Qian, T. Kwon, M. Pollefeys, F. Bogo, S. Tang, Egobody: Human body shape and motion of interacting people from head-mounted devices, in: ECCV, 2022.