RAM: Recover Any 3D Human Motion in-the-Wild

Sen Jia[1,*], Ning Zhu[2,*], Jinqin Zhong[2], Jiale Zhou[3], Huaping Zhang[4],
Jenq-Neng Hwang[1], Lei Li[4,†]

Abstract

RAM incorporates a motion-aware semantic tracker with adaptive Kalman filtering to achieve robust identity association under severe occlusions and dynamic interactions. A memory-augmented Temporal HMR module further enhances human motion reconstruction by injecting spatio-temporal priors for consistent and smooth motion estimation. Moreover, a lightweight Predictor module forecasts future poses to maintain reconstruction continuity, while a gated combiner adaptively fuses reconstructed and predicted features to ensure coherence and robustness. Experiments on in-the-wild multi-person benchmarks such as PoseTrack and 3DPW, demonstrate that RAM substantially outperforms previous state-of-the-art in both Zero-shot tracking stability and 3D accuracy, offering a generalizable paradigm for markerless 3D human motion capture in-the-wild.

1. Introduction

Multiple human 3D motion recovery from monocular videos[8, 13, 14, 18, 37, 38, 43] aims to robustly track and reconstruct temporally coherent 3D human meshes in real time. Accurate 3D motion data plays a vital role in a wide range of applications, including sports analytics, human–computer interaction, medical rehabilitation, and virtual content creation. Traditional motion capture systems, while offering high precision, depend on multi-view camera setups or wearable markers[26, 54], which are expensive, intrusive, and difficult to deploy in real-world environments. In contrast, monocular video provides a low-cost and non-invasive alternative that enables markerless 3D motion estimation from easily captured in-the-wild videos. As a result, achieving robust and accurate multi-person 3D motion recovery from monocular videos has become an important and active area of research[43].

Early approaches such as HMR[20], SPIN[24], and PARE[23] demonstrated single-person 3D mesh reconstruction from static images through end-to-end regression of SMPL[33]. Despite these advances, these methods are tailored for single-person scenarios and cannot effectively handle complex, dynamic scenes with multiple interacting individuals. To overcome this limitation, recent research has shifted toward multi-person settings. 4DHuman[12] combines HMR2.0[12] with PHALP[40]-based tracking to support frame-by-frame mesh reconstruction for multiple subjects, while CoMotion[37] jointly optimizes tracking and modeling within an end-to-end framework. Although these methods mark meaningful progress, multi-person motion recovery in-the-wild remains an unsolved challenge. Most existing methods rely heavily on 2D appearance features and the Hungarian algorithm for identity association[3, 45]. This strategy is highly sensitive to fast motion, severe occlusion, and viewpoint changes, which often result in identity switches or lost tracks. Once identity continuity is broken, the corresponding 3D motion sequence becomes inconsis-

Figure 1. RAM performs online 3D motion reconstruction from monocular video via semantic, motion-aware tracking and occlusion-robust prediction. Unlike prior methods that rely on frame-wise regression and queue-based identity matching, which are often unstable under fast motion and occlusions, RAM maintains consistent identity and accurate reconstruction, while achieving real-time performance that is 2–3× faster than previous approaches. Arrows highlight key differences.

* These authors contributed equally to this work.
† Corresponding Author. (lenny.lilei.cs@gmail.com)
[1]University of Washington [2]Anhui University [3]East China University of Science and Technology [4]Beijing Institute of Technology

tent, producing fragmented or duplicated trajectories that degrade reconstruction accuracy. When targets are partially occluded or undergo fast motion, current methods tend to lose track because they depend solely on visible frame information and lack memory-based motion priors[10, 31, 50]. This leads to discontinuous or unstable reconstructions once the subject reappears. Furthermore, unstable tracking frequently triggers redundant detection, repeated model initialization, and unnecessary computation, preventing real-time performance and scalability. To address these challenges, we propose the Recover-Anyone Module (**RAM**), a unified framework for real-time and robust multi-person 3D motion recovery from monocular videos. RAM integrates three complementary components to jointly improve tracking stability, occlusion handling, and temporal coherence. (1) The SegFollow module introduces **motion-aware priors** via adaptive Kalman filtering to ensure stable identity association under fast motion and heavy occlusion. (2) The Temporal HMR (T-HMR) module enhances 3D reconstruction by injecting **spatio-temporal cues** from adjacent frames, producing smooth and consistent mesh sequences. (3) A lightweight Motion **Predictor** forecasts future poses based on historical motion patterns, maintaining continuity during occluded frames. A **Combiner** then fuses reconstructed and predicted features to achieve coherent long-term motion recovery.

Extensive evaluations on tracking and recovery benchmarks such as PoseTrack [1] and 3DPW [44], we demonstrate that RAM achieves state-of-the-art performance on video-based multi-person 3D motion recovery. These results underscore RAM's potential to advance human-centric AI, providing a robust foundation for advancing research in motion understanding, social interaction, and multi-agent modeling. Our contributions can be summarized as follows:

- We propose **RAM**, a unified framework that combines motion-aware tracking and temporal mesh recovery to achieve robust, coherent, and occlusion-resilient multi-person 3D motion reconstruction.
- RAM achieves state-of-the-art performance in **Zero-shot tracking stability**, reconstruction accuracy, and inference efficiency across standard benchmarks, providing a solid foundation for diverse downstream tasks.
- RAM is the **first** method to achieve **stable, zero-shot multi-person motion recovery** in long, real-world monocular videos with minimal ID switches and without retraining, **bridging the sim-to-real gap** and enabling scalable 3D human capture in-the-wild.

2. Related Works

Tracking and Motion Modeling Recent advances in single-object tracking (SOT) have been driven by end-to-end frameworks and stronger appearance modeling. Early tracking-by-detection pipelines [22] decouple target localization and temporal association, but their loosely coupled design often leads to error accumulation under long-term or occluded scenarios [27, 46, 51, 55]. More recent unified architectures, such as MixFormer [9] and SAM2 [41], integrate target representation and tracking into a single model and even introduce memory mechanisms to enhance temporal stability. While these approaches achieve impressive zero-shot single-target tracking performance, they remain sensitive to heavy occlusions, rapid motion, and large appearance changes, which commonly occur in real-world HOI scenarios. Multi-Object tracking methods follow a tracking-by-detection paradigm, combining object detections with identity association across frames. Classical approaches [40] leverage Kalman filtering [19], Hungarian matching [25], or handcrafted motion cues, while recent works such as Tracktor [2], MotionTrack [39], and MambaTrack [47] adopt learned appearance, motion fusion for improved trajectory consistency [5, 17, 34]. However, MOT systems generally rely on dataset-specific detectors, or domain-specific finetuning. As a result, they remain far from robust in in-the-wild settings with crowding, fast motions, and domain shifts.

Human Motion Reconstruction Human motion estimation has progressed from single-frame, single-person reconstruction to multi-person, video-based human recovery. Early single-frame approaches [4, 36] focus on estimating 3D body pose or mesh from cropped human regions, following either parametric pipelines such as HMR [20] and its successors (e.g., 4D Humans), which regress SMPL parameters [33], or non-parametric methods that directly predict mesh vertices. While effective for isolated subjects, these approaches struggle in crowded scenes. Recent works therefore adopt transformer-based architectures to reason across multiple human instances within an image, improving robustness in multi-person settings [11, 15, 28]. Building on single-frame reconstruction, video-based human motion estimation extends the challenge to temporal association and long-term identity consistency [16, 21, 42]. Tracking-by-detection remains the dominant paradigm, where 2D detections are linked across frames using appearance cues [6, 7, 29]. Systems such as PHALP [40], 4D Humans [12] and CoMotion [37] further incorporate 3D features to enhance association robustness under occlusions, while attention-based matching strategies have been explored to improve identity stability. However, these methods still rely on dataset-specific tuning and motion matching, limiting their generalization and robustness in in-the-wild scenarios.

3. Method

RAM is a unified framework for real-time and accurate multi-person 3D human motion reconstruction from monocular videos. As illustrated in Figure 2, RAM consists of four key components:

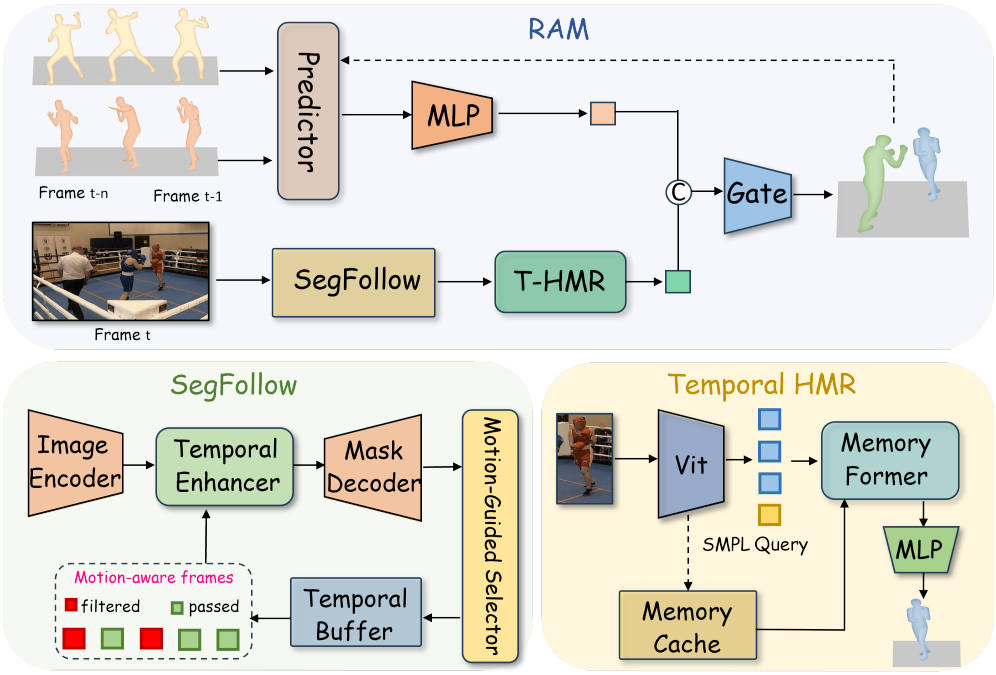

Figure 2. Overview of RAM. The framework integrates four components: SegFollow for motion-guided temporal tracking, Temporal HMR for memory-based 3D reconstruction, a Predictor for motion forecasting under occlusion, and a gated Combiner for robust recovery.

1. **SegFollow** performs motion-aware semantic tracking to maintain stable identity associations across frames, even under occlusion and rapid motion.
2. **T-HMR** reconstructs 3D human meshes from the tracked instances by incorporating temporal context via memory-augmented attention, enabling coherent and robust mesh estimation.
3. **Predictor** models motion dynamics from past reconstructions and forecasts future pose states, offering strong priors when current observations are unreliable.
4. **Combiner** fuses predictions and reconstructions through a learnable gating mechanism, producing stable SMPL outputs with improved temporal consistency under uncertainty.

3.1. SegFollow Module

The SegFollow module builds upon SAM2 to enable robust identity tracking across frames. While SAM2 offers strong segmentation performance, its naive FIFO-based memory update lacks temporal reliability modeling, often leading to identity switches and noise accumulation. SegFollow addresses this by introducing two components: a **motion-guided selector**, which combines Kalman filtering and motion-aware scoring for reliable mask association; and a **temporal buffer**, which updates memory with confidence-weighted smoothing to preserve temporal continuity. Together, they augment SAM2 with explicit motion reasoning and stable long-term tracking.

3.1.1. Motion-Guided Selector

While long-term temporal aggregation is critical for consistent tracking, unreliable associations can corrupt memory with noisy features. To mitigate this, we introduce a motion-guided selector that fuses appearance similarity from SAM2 with motion-aware prior, enabling a more robust and explicit motion reasoning to obtain reliable temporal associations. For a detected instance at time step k, we represent its observation as a bounding box $\mathbf{z}_k = [x_k, y_k, w_k, h_k]^T$ and model the underlying motion state as:

$$\mathbf{x}_k = [x_k, y_k, w_k, h_k, \dot{x}_k, \dot{y}_k, \dot{w}_k, \dot{h}_k]^T.$$

To estimate the motion state of the object, we apply a Kalman Filter (KF), which models temporal dynamics under Gaussian noise and linear assumptions. Given the previous state, the KF predicts a bounding box $\mathbf{H}\hat{\mathbf{x}}_k^-$ for the next frame and estimates its uncertainty. The filter is conditionally updated only when the observation is sufficiently reliable, allowing robust state propagation under partial or noisy observations.

Motion–Aware Selection. For each candidate segmentation mask M_i from SAM2, its bounding box $\mathbf{z}_{k,i}$ is compared to the KF prediction via an IoU-based motion-consistency score:

$$s_{\text{kf}}(M_i) = \text{IoU}\big(\mathbf{H}\hat{\mathbf{x}}_k^-,\, \mathbf{z}_{k,i}\big)\,.$$

We combine this score with SAM2's mask affinity $s_{\text{mask}}(M_i)$ via a gated sum:

$$s_{\text{fused}}(M_i) = \alpha\, s_{\text{mask}}(M_i) + (1-\alpha)\, s_{\text{kf}}(M_i),$$

The mask with the highest s_{fused} is selected. This fusion improves association robustness in challenging scenes, enabling SegFollow to maintain stable tracking even under fast motion or occlusion.

Confidence-Gated Update. To avoid updating the motion state with noisy observations, we adopt a confidence-gated update strategy. Let $s_{\text{obj}}(M^\star)$ denote the confidence score for the chosen mask, generated by the mask decoder. If the confidence exceeds a threshold, we increment a counter C_k that tracks consecutive reliable associations. The KF posterior is updated only when sufficient evidence has accumulated ($C_k \geq \tau_{kf}$); otherwise, the previous state is retained:

$$\hat{\mathbf{x}}_k^+ = \begin{cases} \hat{\mathbf{x}}_k^- + \mathbf{K}_k\big(\mathbf{z}_k - \mathbf{H}\hat{\mathbf{x}}_k^-\big), & C_k \geq \tau_{kf}, \\ \hat{\mathbf{x}}_{k-1}^+, & \text{otherwise.} \end{cases}$$

This gated update prevents unreliable detections, such as during occlusion or rapid motion, from corrupting the motion state, yielding more stable identity tracking compared to SAM2's original FIFO-based memory update.

3.1.2. Temporal Buffer

To further enhance temporal stability, we design a temporal buffer that adaptively updates the memory bank $\mathcal{B}_t$ with smooth weighting. Unlike the FIFO strategy in SAM2, our update mechanism formulates memory update as an exponential moving average, adaptively modulated by motion confidence. At each time step t, the memory embedding $\mathcal{B}_{t-1}$ and the current key feature K_t are combined as:

$$\mathcal{B}_t = \gamma_t \mathcal{B}_{t-1} + (1-\gamma_t) K_t,$$

where $\gamma_t = 1 - \min(s_{\text{kf}}(M^*), \tau_\gamma)$ is an adaptive decay factor modulated by the Kalman-based motion consistency score. The decay factor γ_t adaptively balances current and historical cues: reliable motion prompts stronger updates from present features, while uncertain motion preserves past memory to maintain temporal consistency. This update scheme selectively incorporates high-confidence frames into the memory, guiding attention with temporally reliable context. By incorporating the motion-guided selector and temporal buffer, **SegFollow enables zero-shot, occlusion-robust object tracking**. This lightweight yet effective design provides a solid foundation for stable and accurate 3D reconstruction downstream.

3.2. T-HMR

Given the temporally tracked instances from SegFollow, T-HMR aims to reconstruct 3D human meshes by regressing SMPL parameters from each frame. As using only current-frame features leads to temporal inconsistency and weak robustness under occlusion, T-HMR introduces two core components: **Memory Cache** for selecting informative and reliable temporal features, and the **MemFormer** for injecting these priors into the reconstruction process. Details are provided below.

3.2.1. Memory Cache

To retrieve useful temporal priors, we design a Memory Cache that adaptively selects the top-k relevant frame features from a temporal window centered at the current frame. Specifically, we collect the features encoded by ViT from L adjacent frames, and stack them into a memory feature $F_{\text{mem}} \in \mathbb{R}^{L\times N\times d}$, where N is the number of spatial tokens and d is the feature dimension. We first pool the current frame representations and memory feature across spatial dimensions into global representations $\bar{F}_t$ and $\bar{F}_{\text{mem}} \in \mathbb{R}^{L\times d}$, respectively. Then, we adopt a **dual-branch scoring mechanism** to rank the importance of each frame in F_{mem}. A unified attention-based scoring function is defined as:

$$\mathcal{A}(F_q, F_k) = \text{softmax}_N\left[\frac{(F_q W_q)(F_k W_k)^\top}{\sqrt{d}}\right] \in \mathbb{R}^{N\times N},$$

where $F_q, F_k \in \mathbb{R}^{N\times d}$ and $W_q, W_k \in \mathbb{R}^{d\times d}$. For the **first branch**, we compute the attention scores between the current frame and memory frames to measure relevance and dependencies. For the **second branch**, we evaluate the internal consistency of memory frames through self-attention. The overall importance score is then obtained by combining both branches. Formally:

$$s = \mathcal{A}(\bar{F}_t, \bar{F}_{\text{mem}}) + \text{mean}\left(\mathcal{A}(\bar{F}_{\text{mem}}, \bar{F}_{\text{mem}})\right),$$

Finally, the top-k frames with highest overall importance scores are selected for the memory bank. By modeling both cross-frame relevance and intra-frame consistency, Memory Cache distills high-quality temporal cues while discarding redundant frame features. This facilitates reliable reconstruction under occlusion and motion blur, while ensuring efficiency for MemFormer reasoning.

3.2.2. MemFormer

Given the selected memory features from the Memory Cache and the current frame feature, MemFormer aims to integrate temporal priors for SMPL regression. The overall architecture is composed of N stacked blocks, each following the design described below. We first concatenate a learnable SMPL token with the current frame feature and apply self-attention to model intra-frame interactions. The resulting

representation serves as the query to perform temporal cross-attention, where the memory features are pooled along the spatial dimension to provide motion-consistent keys and values. The output is then passed to a second cross-attention block, which uses memory pooled along the temporal dimension to inject spatially aligned semantics from recent frames. Finally, the updated SMPL token is extracted and decoded via an MLP head to regress the SMPL pose and shape parameters for the current frame. This design allows T-HMR to inject fine-grained temporal priors into the current-frame reconstruction process, enhancing the model's robustness to occlusion and smoothness.

3.3. Predictor

To enhance robustness under occlusion, the Predictor forecasts future motion based on recent reconstruction history. Let $\mathbf{S}_t$ denote the reconstructed motion state at time t, we maintain a FIFO queue $\mathcal{Q}_t = \{\mathbf{S}_{t-T+1}, \ldots, \mathbf{S}_t\}$. This sequence is fed into a stack of L Transformer blocks to capture motion dynamics and predict a latent representation $\hat{Z}t + 1$ for the next frame. The output latent $\hat{Z}_{t+1}$ is then passed to the Combiner module as a motion-conditioned prior, supporting stable reconstruction when current-frame cues are unreliable. The motion queue is updated online during inference, enabling real-time adaptation to dynamic motion patterns.

3.4. Combiner

The Combiner integrates the motion prior from the Predictor and the current reconstruction feature from T-HMR to produce a stable next-step motion. It fuses the representations via a learnable gate and regresses SMPL parameters in a single head. Given the T-HMR feature Z^{h}_{t+1} of the current frame and the predicted latent motion prior $\hat{Z}_{t+1}$ from the Predictor, the Combiner predicts a gating vector via an MLP layer:

$$g_{t+1} = \sigma\Big(\mathrm{MLP}_g\Big([Z^{\mathrm{h}}_{t+1}, \hat{Z}_{t+1}]\Big)\Big) \in [0,1]^d,$$

where $\sigma(\cdot)$ is the sigmoid function. The fused feature is then computed using a weighted interpolation:

$$Z^{\mathrm{c}}_{t+1} = (1 - g_{t+1}) \odot Z^{\mathrm{h}}_{t+1} + g_{t+1} \odot \hat{Z}_{t+1},$$

where $\odot$ denotes element-wise multiplication. This design encourages reliance on T-HMR features under confident observations, while shifting toward the predicted prior under occlusion or uncertainty, enabling stable and consistent motion recovery. Finally, a regression head maps Z^{c}_{t+1} to SMPL parameters for next-frame reconstruction.

3.5. Training Objectives

We adopt a three-stage learning strategy to train RAM. **Stage1**. Following 4D-Humans, we pretrain T-HMR's image encoder and pose regression module on large batches of single images. The SMPL supervision employs multiple objectives: L1 loss on 2D joint projections and pelvis-relative 3D joint positions, L2 loss on joint rotations, L1 loss on shape parameters (betas, for fully annotated samples), binary cross-entropy loss for prediction confidence (with unmatched outputs as negatives), and a keypoint heatmap loss for spatial consistency.

Stage2. We train the temporal predictor with a scheduled sampling strategy exclusively on 8-frame 3D pose sequences and synthetic 3D pose sequences generated by panning/zooming single-image pose data from InstaVariety. The 3D and 2D pose points of the first frame are matched to high-quality ground-truth annotations, and the predictor is unrolled over time to perform temporal prediction of pose states up to the n-th frame. Each timestep is supervised using a multi-term loss that jointly optimizes 3D pose point accuracy and 2D pose point projection consistency (consistent with the SMPL-based loss framework of Stage 1).

Stage3. With the Predictor and RAM components frozen, we fine-tune the full framework to learn dynamic fusion under occlusion. We simulate occlusion by randomly masking 60% of human body regions in the input images, encouraging the Combiner to rely more on the predicted motion prior when visual cues are incomplete. This strategy promotes robust recovery by guiding the model to adaptively balance reconstruction and prediction. Supervision follows the Stage-1 SMPL loss.

4. Experment

To evaluate the effectiveness of RAM, we conducted comprehensive experiments on multiple datasets, assessing both its tracking and estimation capabilities. Our evaluation includes standard benchmarks as well as challenging scenarios characterized by rapid motion, frequent occlusions, and the presence of multiple interacting subjects such as in basketball and boxing scenes.

4.1. Implementation Details

4.1.1. Evaluation Datasets

We evaluate RAM across multiple benchmarks covering 2D/3D pose estimation and multi-person tracking. For tracking and 2D keypoint estimation, we use the PoseTrack [1] and COCO dataset [32], which feature frequent occlusions, complex motions and diverse poses. In addition, TrackID-3x3 [48] and Olympic Boxing dataset further challenge robustness in real-world sports with dense interactions and fast motion, serving as rigorous benchmarks for evaluating generalization and robustness. For 3D reconstruction, we adopt 3DPW [44], which offers accurate mesh and joint annotations in unconstrained outdoor settings.

Table 1. Comparison on PoseTrack18 and PoseTrack21: while 4DHumans and CoMotion are trained on these datasets, RAM is evaluated zero-shot without retraining, and achieves the best results across all metrics.

Method	PoseTrack18			PoseTrack21						
	HOTA↑	IDs↓	MOTA↑	MOTA↑	IDF1↑	IDP↑	IDR↑	FP↓	FN↓	FPS↑
4DHumans [12]	57.8	382	61.4	–	–	–	–	–	–	-
4DHumans (reproduced)*	58.0	349	**61.8**	56.7	70.9	87.1	59.7	7817	50652	0.51
CoMotion [37] *strict*	58.2	232	59.9	61.8	74.0	89.1	63.3	6086	45664	-
CoMotion [37]	54.9	344	51.3	71.4	79.5	87.1	73.0	8115	30394	5.68
RAM (**Zero-shot**)	**66.4**	**15**	57.7	**74.4**	**85.9**	**93.8**	**79.2**	**5864**	**24044**	**10.32**

Table 2. Comparison on two challenging real-world sports scenarios featuring frequent occlusions, and fast motion. We evaluate zero-shot generalization of prior methods and RAM, and additionally report an ablation using only SAM2-based tracking.

Method	TrackID3x3 (Indoor)			TrackID3x3 (Outdoor)		
	TI-HOTA↑	TI-DetA↑	TI-AssA↑	TI-HOTA↑	TI-DetA↑	TI-AssA↑
4DHumans [12]	5.17	4.89	5.47	2.66	2.19	3.22
CoMotion [37]	42.20	32.93	54.07	30.87	23.06	41.33
RAM (SAM2)	54.23	48.39	60.78	38.60	36.92	40.37
RAM	**75.07**	**62.87**	**89.66**	**66.68**	**51.39**	**86.63**

4.1.2. Evaluation Metrics

Following the evaluation protocol of previous works [37], we report standard metrics including Multiple Object Tracking Accuracy (MOTA), Identity F1 Score (IDF1), number of ID switches (IDs), ID precision and recall (IDP, IDR), and Hybrid Object Tracking Accuracy (HOTA). We also include error statistics such as false positives and false negatives for comprehensive analysis. Among these, MOTA primarily evaluates the completeness of detection, whereas IDF1, IDs, IDP, and IDR quantify how well the tracker maintains identity consistency over time. For TrackID-3x3, we additionally report TI-HOTA, TI-DetA, and TI-AssA [48], which evaluate ID-consistent tracking by disentangling detection and association accuracy. For 2D pose estimation on COCO [32] and PoseTrack [1], we use the Percentage of Correct Keypoints (PCK), which measures the proportion of keypoints predicted within a normalized distance threshold. For 3D pose evaluation on 3DPW [44], we report the Mean Per Joint Position Error (MPJPE) and its Procrustes-aligned variant (PA-MPJPE), which account for absolute and rigid-aligned joint accuracy, respectively.

4.2. Results

4.2.1. Tracking Results

We compare against state-of-the-art 3D recovery methods including 4DHumans and CoMotion. While other multi-object tracking methods exist [35, 49], they are not tailored for 3D recovery and typically require tuning for domain transfer. For completeness, we include comparisons in the supplementary, where our SegFollow still achieves superior zero-shot tracking ability.

Evaluation on PoseTrack We evaluate RAM on the PoseTrack benchmarks to assess its tracking performance under challenging scenarios involving occlusion, fast motion, and frequent identity interactions. As shown in Table 1, RAM consistently outperforms prior state-of-the-art methods such as 4DHumans and CoMotion across key dimensions, including identity consistency, multi-person detection coverage, and robustness. Due to their reliance on 3D trajectory matching, both 4DHumans and CoMotion exhibit poor performance under occlusion or viewpoint shifts, often resulting in identity switches and fragmented tracks.

In contrast, RAM attains a HOTA of 66.4 with merely **15 ID switches** on PoseTrack18, marking an **order-of-magnitude** gain in identity stability over prior methods; the slight drop in MOTA stems largely from stricter association behavior rather than tracking failures. On PoseTrack21, RAM sets a new benchmark with 74.4 MOTA, and further achieves **+6.4** IDF1 and **+82%** FPS improvements over CoMotion, establishing RAM as the **new state-of-the-art for real-time, stable human motion tracking** in-the-wild. These gains stem from the SegFollow module, which integrates motion-aware priors via a Kalman-based selector for robust mask association, and employs a temporal buffer that focuses historically reliable frames, ensuring stable identity propagation even under occlusion and motion blur.

These results highlight RAM's effectiveness in enabling

Table 3. **Pose estimation.** Normalized PCK accuracy on projected 2D keypoints at varying thresholds on the COCO and PoseTrack datasets, alongside MPJPE of 3D keypoints on the 3DPW dataset. We highlight that our model performs similarly when provided the full image as input rather than an oracle-resized crop around a target person. See text for analysis.

Table 4. Your caption here (e.g., Quantitative comparison on 2D and 3D pose estimation benchmarks).

	Method	COCO		PoseTrack		3DPW	
		PCKn@0.05↑	PCKn@0.1↑	PCKn@0.05↑	PCKn@0.1↑	MPJPE↓	PA-MPJPE↓
2D/3D HPE	PyMAF [52]	0.68	0.86	0.77	0.92	92.8	58.9
	CLIFF [30]	0.64	0.88	0.75	0.92	69.0	43.0
	PARE [23]	0.72	0.91	0.79	0.93	82.0	50.9
	PyMAF-X [53]	0.79	0.93	0.85	0.95	78.0	47.1
	HMR 2.0a [12]	0.79	0.95	0.86	0.97	70.0	44.5
	HMR 2.0b [12]	0.86	0.96	0.90	**0.98**	81.3	54.3
	Comotion [37]	0.79	0.92	0.88	0.96	60.0	37.3
	RAM (Ours)	**0.89**	**0.97**	**0.93**	**0.98**	**53.0**	**34.1**

robust and consistent multi-person tracking in wild, real-world video scenarios. Notably, **RAM achieves this impressive performance in a zero-shot setting, without retraining on PoseTrack**, highlighting its remarkable generalization to in-the-wild tracking scenarios.

Zero-Shot Evaluation on Challenging Real-World Videos We evaluate RAM's zero-shot generalization on TrackID-3x3 [48], a challenging basketball competition dataset from real-world featuring frequent occlusions and fast motion. Compared to PoseTrack's short clips (∼3s), TrackID-3x3 includes longer videos (∼10s indoor, ∼40s outdoor), making the outdoor setting more challenging due to video length and in-the-wild complexity.

As shown in Table 2, RAM significantly outperforms prior methods, achieving **+78%** (indoor) and **+116%** (outdoor) higher TI-HOTA than CoMotion. These gains highlight RAM's robustness in real-world scenarios with frequent occlusion and multi-person interaction. In contrast, 4DHumans and CoMotion underperform due to their reliance on domain-specific trajectory matching.

Our **ablation with SAM2**-based tracking alone results in notable performance drops, particularly in outdoor scenarios, despite using a strong pre-trained segmenter. While SAM2 provides semantic initialization, it lacks motion priors and temporal modeling, leading to unstable tracking under occlusions and fast motion. Its TI-HOTA on outdoor TrackID-3x3 only improves over CoMotion by **+7.7**, indicating limited generalization. In contrast, full RAM yields a **+35.8** gain, demonstrating that our **motion-aware designs are key** to achieving robust, real-world tracking and cannot be replaced by strong segmentation alone.

Together, these results establish RAM as the first framework to robustly generalize to long, in-the-wild multi-human videos under zero-shot settings.

4.2.2. Estimation Results

Evaluation on PoseTrack and COCO We evaluate RAM on the PoseTrack and COCO datasets to assess its reconstruction accuracy in real-world complex scenarios. These 2D datasets encompass diverse human motions and challenging real-world conditions such as crowd interactions and occlusion, providing a rigorous benchmark for evaluating model robustness and generalization. As shown in Table 3, RAM achieves state-of-the-art PCK accuracy across both datasets, outperforming prior works including PARE, CLIFF, and HMR 2.0. Notably, RAM achieves **0.93** PCK@0.05 on PoseTrack and **0.89** on COCO, indicating strong localization precision even under dense, multi-person settings. Compared to CoMotion, which already integrates temporal cues, RAM still yields consistent gains, highlighting its temporal stability and occlusion robustness.

These gains primarily derive from T-HMR's temporal priors and the Predictor's motion-conditioned inference, which together enable robust estimation under occlusion. The Combiner further refines this by adaptively balancing predicted and observed cues, ensuring consistent recovery even when joints are fully invisible. These results highlight RAM's effectiveness in delivering stable and accurate pose estimation across complex, real-world conditions.

Evaluation on 3DPW We further evaluate RAM on the 3DPW dataset to assess its 3D recovery accuracy. Following the CoMotion protocol, we report standard 3D metrics including MPJPE and PA-MPJPE, as shown in Table 3. RAM achieves the lowest reconstruction error among all methods, with **53.0** MPJPE and **34.1** PA-MPJPE, outperforming CoMotion and recent single-person models such as HMR 2.0. The improvement primarily stems from T-HMR's temporal

Figure 3. Qualitative comparison on the Olympics Boxing dataset. Both 4DHumans and CoMotion suffer from identity switches and tracking failures under fast motion and heavy occlusion, resulting in fragmented 3D reconstructions and repeated identity reinitialization. This not only degrades reconstruction quality but also leads to high inference overhead. In contrast, RAM robustly tracks boxers and the referee throughout the sequence with consistent identity association and real-time, accurate 3D motion recovery.

priors, which strengthen frame-to-frame consistency. In addition, the synergy between current-frame reconstruction and predictive motion priors further enhances robustness. These results demonstrate that RAM is capable of performing accurate 3D human mesh reconstruction directly from monocular videos, enabling robust and coherent multi-person motion recovery in-the-wild.

4.2.3. Qualitative Results

We evaluate RAM on two real-world sports datasets with fast motion, occlusion, and complex multi-person interaction. As shown in Fig. 2 and Fig. 3, existing methods such as 4DHumans and CoMotion struggle to maintain identity consistency (Person colors are used to denote their tracking IDs) and recover accurate 3D motion. CoMotion often fails under occlusion or motion blur, while 4DHumans exhibits frequent identity switches and unstable queue management, resulting in poor reconstruction and high latency. These limitations prevent existing methods from achieving robust zero-shot multi-human motion recovery in real-world settings.

In contrast, RAM achieves real-time, robust tracking and accurate 3D recovery throughout. Its motion-aware tracking and occlusion-resilient prediction allow it to sustain identity and reconstruction quality even in long and dynamic videos such as boxing sequences.

These results demonstrate RAM's strong **zero-shot generalization** and its **practical applicability** for 3D motion recovery in unconstrained real-world settings.

5. Conclusion

We propose RAM, a unified framework for real-time and robust multi-person 3D motion reconstruction from monocular videos. By integrating semantic tracking with motion-aware modeling, RAM effectively mitigates ID switches and tracking loss under occlusion and viewpoint changes, enabling stable identity association and reducing redundant computation. Leveraging temporal priors and motion prediction, RAM achieves smooth and accurate reconstruction even under severe occlusions. Extensive experiments demonstrate that RAM consistently outperforms prior works, especially in complex real-world sports scenes, achieving substantial gains in tracking stability, reconstruction quality, and computational efficiency. RAM provides a solid foundation for future research in human-centric motion understanding and will be extended to human-object recovery.

6. Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFC3308101; in part by the National Natural Science Foundation of China under Grant 62401005; and in part by

the Natural Science Foundation of Anhui Higher Education Institutions of China under Grant 2023AH050069.

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018.

[2] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 941–951, 2019.

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. Ieee, 2016.

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016.

[5] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6158–6166, 2017.

[6] Mohamed Chaabane, Peter Zhang, J Ross Beveridge, and Stephen O'Hara. Deft: Detection embeddings for tracking. *arXiv preprint arXiv:2102.02267*, 2021.

[7] Zhiwei Chen, Yupeng Hu, Zhiheng Fu, Zixu Li, Jiale Huang, Qinlei Huang, and Yinwei Wei. Intent: Invariance and discrimination-aware noise mitigation for robust composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[8] Yu Cheng, Bo Wang, Bo Yang, and Robby T Tan. Graph and temporal convolutional networks for 3d multi-person pose estimation in monocular videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1157–1165, 2021.

[9] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13608–13618, 2022.

[10] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10061–10072, 2023.

[11] Yuan Dong, Chuan Fang, Liefeng Bo, Zilong Dong, and Ping Tan. Panocontext-former: Panoramic total scene understanding with a transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28087–28097, 2024.

[12] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023.

[13] Brian Gordon, Sigal Raab, Guy Azov, Raja Giryes, and Daniel Cohen-Or. Flex: extrinsic parameters-free multi-view 3d human motion reconstruction. In *European Conference on Computer Vision*, pages 176–196. Springer, 2022.

[14] Ruocheng Gu, Sen Jia, Yule Ma, Jinqin Zhong, Jenq-Neng Hwang, and Lei Li. Mocount: Motion-based repetitive action counting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9026–9034, 2025.

[15] Markus Hiller, Krista A Ehinger, and Tom Drummond. Perceiving longer sequences with bi-directional cross-attention transformers. *Advances in Neural Information Processing Systems*, 37:94097–94129, 2024.

[16] Tao Hu, Liwei Wang, Xiaogang Xu, Shu Liu, and Jiaya Jia. Self-supervised 3d mesh reconstruction from single images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6002–6011, 2021.

[17] Yupeng Hu, Zixu Li, Zhiwei Chen, Qinlei Huang, Zhiheng Fu, Mingzhu Xu, and Liqiang Nie. Refine: Composed video retrieval via shared and differential semantics enhancement. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2026.

[18] Zeren Jiang, Chen Guo, Manuel Kaufmann, Tianjian Jiang, Julien Valentin, Otmar Hilliges, and Jie Song. Multiply: Reconstruction of multiple people from monocular video in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 109–118, 2024.

[19] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

[20] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018.

[21] Hossein Feizollah Zadeh Khoiee, David Labbe, Thomas Romeas, Jocelyn Faubert, and Sheldon Andrews. Multi-person physics-based pose estimation for combat sports. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5832–5841, 2025.

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[23] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11127–11137, 2021.

[24] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. 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*, pages 2252–2261, 2019.

[25] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[26] Jiye Lee and Hanbyul Joo. Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1091–1100, 2024.

[27] Lei Li, Sen Jia, Jianhao Wang, Zhongyu Jiang, Feng Zhou, Ju Dai, Tianfang Zhang, Zongkai Wu, and Jenq-Neng Hwang. Human Motion Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[28] Lei Li, Sen Jia, and Jenq-Neng Hwang. Multiple human motion understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6297–6305, 2026.

[29] Peixuan Li and Jieyu Jin. Time3d: End-to-end joint monocular 3d object detection and tracking for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3885–3894, 2022.

[30] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022.

[31] Zixu Li, Yupeng Hu, Zhiwei Chen, Qinlei Huang, Guozhi Qiu, Zhiheng Fu, and Meng Liu. Retrack: Evidence driven dual stream directional anchor calibration network for composed video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

[34] Juanwu Lu, Wei Zhan, Masayoshi Tomizuka, and Yeping Hu. Towards generalizable and interpretable motion prediction: A deep variational bayes approach. In *International Conference on Artificial Intelligence and Statistics*, pages 4717–4725. PMLR, 2024.

[35] Weiyi Lv, Yuhang Huang, Ning Zhang, Ruei-Sung Lin, Mei Han, and Dan Zeng. Diffmot: A real-time diffusion-based multiple object tracker with non-linear prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19321–19330, 2024.

[36] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2823–2832, 2017.

[37] Alejandro Newell, Peiyun Hu, Lahav Lipson, Stephan R Richter, and Vladlen Koltun. Comotion: Concurrent multi-person 3d motion. *arXiv preprint arXiv:2504.12186*, 2025.

[38] Sungchan Park, Eunyi You, Inhoe Lee, and Joonseok Lee. Towards robust and smooth 3d multi-person pose estimation from monocular videos in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14772–14782, 2023.

[39] Zheng Qin, Sanping Zhou, Le Wang, Jinghai Duan, Gang Hua, and Wei Tang. Motiontrack: Learning robust short-term and long-term motions for multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17939–17948, 2023.

[40] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3d appearance, location and pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2740–2749, 2022.

[41] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[42] Xiaolong Shen, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Global-to-local modeling for video-based 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8887–8896, 2023.

[43] Nicolas Ugrinovic, Boxiao Pan, Georgios Pavlakos, Despoina Paschalidou, Bokui Shen, Jordi Sanchez-Riera, Francesc Moreno-Noguer, and Leonidas Guibas. Multiphys: Multi-person physics-aware 3d motion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2331–2340, 2024.

[44] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, 2018.

[45] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[46] Zijie Wu, Chaohui Yu, Yanqin Jiang, Chenjie Cao, Fan Wang, and Xiang Bai. Sc4d: Sparse-controlled video-to-4d generation and motion transfer. In *European Conference on Computer Vision*, pages 361–379. Springer, 2024.

[47] Changcheng Xiao, Qiong Cao, Zhigang Luo, and Long Lan. Mambatrack: a simple baseline for multiple object tracking with state space model. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 4082–4091, 2024.

[48] Kazuhiro Yamada, Li Yin, Qingrui Hu, Ning Ding, Shunsuke Iwashita, Jun Ichikawa, Kiwamu Kotani, Calvin Yeung, and Keisuke Fujii. Trackid3x3: A dataset and algorithm for multi-player tracking with identification and pose estimation in 3x3 basketball full-court videos. In *Proceedings of the 8th International ACM Workshop on Multimedia Content Analysis in Sports*, pages 163–173, 2025.

[49] Mingzhan Yang, Guangxin Han, Bin Yan, Wenhua Zhang, Jinqing Qi, Huchuan Lu, and Dong Wang. Hybrid-sort: Weak cues matter for online multi-object tracking. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6504–6512, 2024.

[50] En Yu, Zhuoling Li, and Shoudong Han. Towards discriminative representation: Multi-view trajectory contrastive learning for online multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8834–8843, 2022.

[51] En Yu, Zhuoling Li, Shoudong Han, and Hongwei Wang. Relationtrack: Relation-aware multiple object tracking with decoupled representation. *IEEE Transactions on Multimedia*, 25:2686–2697, 2022.

[52] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11446–11456, 2021.

[53] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12287–12303, 2023.

[54] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4d association graph for realtime multi-person motion capture using multiple video cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1324–1333, 2020.

[55] Chong Zhou, Chenchen Zhu, Yunyang Xiong, Saksham Suri, Fanyi Xiao, Lemeng Wu, Raghuraman Krishnamoorthi, Bo Dai, Chen Change Loy, Vikas Chandra, et al. Edgetam: On-device track anything model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13832–13842, 2025.