DETRAM: End-to-end DEtection, Tracking and Recovery of HumAn Meshes

Chunggi Lee[1], Seonwook Park[2], Wanhua Li[3], Umar Iqbal[2,*], and Hanspeter Pfister[1,*]

[1]Harvard University [2]NVIDIA [3]Nanyang Technological University

Fig. 1: Comparison between traditional multi-stage pipelines and our unified one-stage DETRAM framework. Conventional approaches decompose the problem into sequential detection, tracking, and human mesh recovery (HMR) modules, causing error accumulation, increased latency, and difficulty maintaining consistent identities over time. In contrast, DETRAM performs all three tasks jointly in a single feed-forward pass. This enables efficient, end-to-end multi-person reconstruction and tracking without requiring either an external detector or a separate learned tracking module. DETRAM discovers and tracks new people as they enter the scene, while also supporting optional user-prompted tracking, allowing individuals to be added or specified when desired.

Abstract. In the task of human mesh recovery (HMR), multi-person scenes are particularly difficult to handle due to the many entities that appear and occlusions between them over time. In particular for video inputs, there is a need to track each entity reliably and consistently. Existing methods rely on pretrained human detection modules, increasing their runtime and limiting the number of tracked entities. We present DETRAM, a unified framework for multi-person HMR and tracking that simultaneously detects, reconstructs, and tracks humans across time, both automatically and via user prompts. DETRAM uses a single transformer decoder with an identity-consistent set of learnable query embeddings that persist across frames: detection queries discover new people, tracking queries maintain pose and shape for existing individuals, and prompt queries follow user-specified identities. Our approach achieves state-of-the-art tracking results on PoseTrack21, 3DPW, BEDLAM, and MuPoTS-3D, and competitive reconstruction accuracy on BEDLAM and 3DPW, while uniquely supporting prompt-based tracking of individuals in multi-person scenes. To our knowledge, this is the first method to unify promptability and multi-person HMR with tracking in an end-to-end trainable framework, enabling user-directed human analysis in videos.

* Corresponding authors.

1 Introduction

Videos of public places, sporting events, and urban environments typically contain many humans and their movements over time. Tracking each human and recovering their 3D pose and shape in such crowded videos enables applications in sports analytics, mixed reality, safety monitoring, and human behavior understanding. However, these scenes are highly challenging: people frequently occlude one another, body parts are truncated by crops or the camera view, and individuals can enter or leave the scene at any time. These factors often lead to tracking failures or inconsistent 3D reconstructions, especially in long sequences.

Existing methods for multi-person human mesh recovery (HMR) and tracking are mostly built as multi-stage pipelines. A detector first outputs bounding boxes, which are subsequently cropped and fed into pose or mesh recovery networks. Tracking IDs are then assigned in a separate association stage based on appearance or pose features [14,40,41]. While effective in some settings, this design has several weaknesses. Bounding box errors or misaligned crops can remove important spatial context, leading to missing limbs or ambiguous associations in crowded scenes. Moreover, errors propagate across stages, each module may need to be trained independently, and the overall runtime increases.

Recent transformer-based approaches address some of these issues by operating directly on the full image and regressing human pose and shape without cropping, within detection-transformer architectures for multi-person HMR [2,44,45]. By reasoning over all humans jointly via attention, these models better preserve inter-human relationships and global context, and they can deliver more accurate multi-person reconstructions in a single frame. However, they still treat frames independently, with no explicit mechanism for maintaining identity over time or interacting with user-specified individuals. Extending them to video, therefore, requires an external tracking module or post-hoc association strategy, reintroducing the multi-stage issues described above and making it non-trivial to deploy them directly in streaming video settings.

In this work, we take a unified approach to multi-person HMR and tracking in videos by proposing an "End-to-end **DE**tection, **T**racking and **R**ecovery of Hum**A**n **M**eshes" Transformer (DETRAM) architecture. Instead of relying on external detectors or separately trained tracking modules, we define an identity-consistent set of learnable query embeddings that persist across time. Within a transformer-based decoder in line with recent detector-transformer designs [22, 27], these queries take on three complementary roles: (1) *detection queries* discover new humans entering the scene; (2) *tracking queries* carry pose, shape, and identity information forward across frames; and (3) *prompt queries* are initialized from user-specified spatial prompts (i.e. bounding boxes) and enforce controllable tracking of particular individuals throughout a sequence. By jointly decoding all query types in a single feed-forward pass per frame, our model simultaneously detects, reconstructs, and tracks all relevant humans without any explicit cropping or any separately learned tracking stage (as illustrated in Fig. 1).

DETRAM's unified formulation improves robustness to occlusions and crowded scenes, and substantially reduces inference latency compared to multi-stage pipelines. Concurrent query-based methods for multi-person mesh recovery [26,36] also aim to do end-to-end tracking and reconstruction, but either use clip-level motion queries with a separate learnable temporal module [26] or decouple per-frame detection and pose updates into distinct components [36]. In contrast, we use a single framewise decoder in which detection, tracking, and prompt queries are decoded jointly, and only our formulation exposes user-specified prompt queries for controllable tracking. Our approach attains state-of-the-art tracking performance on PoseTrack21, MuPoTS-3D, BEDLAM, and 3DPW, and reconstruction accuracy competitive with or exceeding the best prior methods on BEDLAM and 3DPW, and, to our knowledge, is the first to unify promptability with multi-person HMR and tracking in an end-to-end trainable framework.

In summary, our contributions are: (a) we introduce DETRAM, a single-stage architecture that detects, tracks, and recovers 3D meshes of multiple humans from video, (b) this is enabled by a query design involving identity-consistent tracking queries propagated across frames and prompt queries initialized from user inputs; and (c) experimental results on the PoseTrack21, MuPoTS-3D, 3DPW and BEDLAM datasets show competitive tracking and reconstruction performance.

2 Related Work

Human mesh recovery. Early HMR works from monocular images focused on single-person setups and extended to multiple people using scene-level constraints or multi-stage pipelines [11,16–21,23,24,42,48,52]. Typical systems run a detector, crop each person, and estimate pose and shape independently. More recent approaches use one-stage, transformer-based models that operate on full images and regress meshes for all visible people jointly [2,44–46,48,49], which improves robustness in crowded scenes by keeping global context and modeling inter-person interactions. These methods, however, are still single-frame: they do not maintain identities over time and they offer no mechanism for selecting which individuals to follow in a video. Our work follows the one-stage, full-frame formulation but treats detection, reconstruction, and identity as a single video-level problem by making a subset of queries persistent across frames.

Tracking and reconstructing humans from video. Several methods combine tracking and 3D reconstruction by lifting 2D detections into a 3D representation and tracking in that space. T3DP [40], PHALP [41], and 4DHumans [14] estimate 3D pose and position per person and then associate tracks in this feature space, while keeping detection, HMR, and tracking as separate modules. More recent frameworks move towards end-to-end multi-person motion capture. Motions-as-Queries (MaQ) [26] uses temporal motion queries over clips and predicts holistic whole-body motion directly from video in an offline manner. CoMotion [36] combines a strong per-frame detector with a learned

pose-update network that updates 3D poses online in crowded scenes. These concurrent methods also adopt query or update-based formulations for multi-person mesh recovery, but do not allow user-specified prompts. Our method has a similar goal, multi-person 3D reconstruction with stably tracked identities, but uses a single DETR-style decoder with explicit detection, tracking, and prompt queries, enabling user-directed tracking. The concurrent work, SAM 3 [6], also supports promptable video segmentation and tracking, but focuses on category or exemplar-level 2D mask propagation. In contrast, DETRAM detects, reconstructs, and tracks 3D SMPL parameters via a unified decoder, by treating prompts as identity anchors.

DETR-style set prediction and query design. DETR and follow-up work cast perception as set prediction with learned queries and bipartite matching [7,22,27]. Variants modify query parameterization (e.g., dynamic anchor-box queries in DAB-DETR [27]) and training (e.g., denoising in DN-DETR [22]) to stabilize and speed up learning. Recent multi-person HMR models adopt DETR-style decoders to regress 3D meshes from single images [2, 44, 45], but treat all queries as frame-local. We build on the same set-prediction method, but extend it temporally and structurally: a subset of queries is kept identity-consistent across frames, and explicit prompt queries are added for user control. All query types are optimized jointly with a denoising-style objective, yielding an end-to-end architecture that handles detection, tracking, and mesh recovery of multiple humans in video within a single decoder.

3 Method

Preliminaries. Given an input image sequence $\{I^1, I^2, \ldots, I^T\} \in \mathbb{R}^{T\times h\times w\times 3}$ containing an unknown number of people N_p, our goal is to recover articulated 3D body meshes with temporally consistent identities. We adopt the SMPL [28] parametric model to represent each human with pose, shape, and global translation in the camera coordinate system. For each person i, the mesh with $N_v = 6890$ vertices is defined as, $H_i = \text{SMPL}(\theta_i, \beta_i, \tau_i)$, where $\theta_i \in \mathbb{R}^{J\times 3}$ encodes the $J = 24$ joint rotations (including global orientation), $\beta_i \in \mathbb{R}^{10}$ denotes the low-dimensional shape parameters, and $\tau_i \in \mathbb{R}^3$ is the 3D translation in the camera frame.

Overview. We build upon recent DETR-style architectures [7, 8], extending them to multi-person human mesh reconstruction and identity tracking. As illustrated in Fig. 2, our framework is comprised of an image encoder, a DETRAM decoder, and multiple prediction heads responsible for estimating confidence scores, bounding boxes, and SMPL parameters.

For each frame, the image is divided into fixed-size patches of dimension $P \times P$. Each patch is embedded into a feature token using patch embeddings and positional encodings, producing a token set $\mathcal{T} = \{t_1, t_2, \ldots, t_k\}$, where $k = (h/P) \times (w/P)$. These tokens are processed by a transformer encoder to capture spatial structure across the image. The decoder then operates over a set

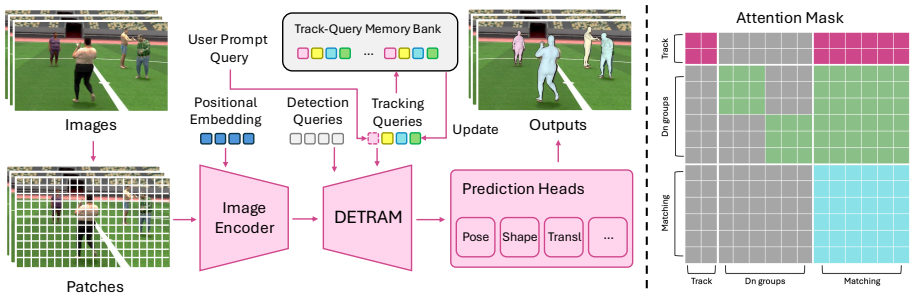

Fig. 2: Overview of the DETRAM framework. Our model (left) encodes input images into visual tokens and jointly processes detection queries (discovering new individuals), tracking queries (identity-persistent embeddings propagated across frames), and optional prompt queries, continuously updated via a track-query memory bank for long-range temporal consistency. Per-person predictions for pose, shape, translation, and confidence are produced through multi-headed regressors. The attention-masking scheme (right) separates Hungarian-matching, denoising, and tracking/prompt queries to prevent denoising GT leakage, allowing DETRAM to learn detection, reconstruction, and identity tracking in a single end-to-end pass while preserving the distinct role of each query type.

of learnable queries $\mathcal{Q} = \{q_1, q_2, \ldots, q_{N_q}\}$, with N_q selected to accommodate the maximum number of people expected in a scene. Through cross-attention with the encoded features, each query progressively evolves into a latent representation corresponding to an individual human instance.

Decoder outputs are passed through lightweight MLP-based prediction heads, including a confidence head $\mathcal{H}_c$, pose head $\mathcal{H}_p$, shape head $\mathcal{H}_s$, translation head $\mathcal{H}_t$, and bounding box head $\mathcal{H}_b$. These heads jointly regress SMPL parameters (pose θ, shape β, translation t) along with 2D bounding boxes. A separate linear projection predicts per-query confidence scores, and only predictions with confidence above a threshold α_d are retained.

3.1 Detection and Tracking Queries.

We tokenize each input image using a DINOv2-based ViT encoder [37]: $F = \text{Encoder}(I)$ with an input resolution of 1288×1288, producing dense visual tokens that serve as the basis for all subsequent query reasoning.

Our formulation unifies detection and tracking directly within the transformer decoding process through three complementary query types: *detection queries*, *tracking queries*, and *prompt queries*. Together, these enable our model to simultaneously discover new people, maintain persistent identities, and integrate user-directed prompts in a single end-to-end framework. All three query types share an anchor-based representation: learned for detection, propagated for tracking, and user-provided for prompts. Each query is represented by an anchor A_q, a content embedding C_q, and an anchor-derived positional embedding. Detection queries are learned, while tracking and prompt queries use propagated or user-provided anchors with TMCA-refreshed contents.

Detection Queries. Detection queries discover and localize human instances in each frame. Instead of relying on preset anchor shapes or external detectors, we *learn* a set of query anchors. Each detection query is parameterized by an anchor box as $A_q = (x_q, y_q, w_q, h_q)$, where (x_q, y_q) denote the normalized center coordinates and (w_q, h_q) the normalized width and height. Each anchor A_q is paired with a content embedding $C_q \in \mathbb{R}^D$ and a positional embedding $P_q \in \mathbb{R}^D$. Through cross-attention with visual tokens, these detection queries identify potential human instances and predict their 2D locations, 3D pose, shape, and confidence, acting as the entry point for newly appearing people.

Tracking and Prompt Queries. In dynamic scenes: as people move, enter or exit the frame, or undergo significant appearance changes due to pose variation, the correspondence between detection queries and human entities drifts, leading to fragmented trajectories and identity errors. To address this, we design our tracking queries to be *identity-persistent*. At every new frame, each tracking query attends to the detection queries and updates its representation through *Track-Modulated Cross-Attention*. This enables the query to absorb relevant spatial context while preserving its identity embedding, ensuring robust temporal consistency even under occlusions or appearance changes. The same mechanism naturally extends to user-driven *prompt queries*, allowing explicit control over which individuals to follow throughout the sequence.

Track-Modulated Cross-Attention. To maintain identity consistency across frames, each tracking query updates its content embedding by attending to the detection queries. Here, N_t denotes the number of tracking queries and N_d denotes the number of detection queries. Given the set of tracking-query anchors from the previous frame $A^{\text{trk}} = \{A_q^{\text{trk}}\}_{q=1}^{N_t}$ and the detection-query anchors $A^{\text{det}} = \{A_k^{\text{det}}\}_{k=1}^{N_d}$, we compute cross-attention weights as

$$w_{qk} = \text{Softmax}_k\big(\phi(A_q^{\text{trk}})^\top \psi(A_k^{\text{det}})\big)\,,$$

where $\phi(\cdot)$ and $\psi(\cdot)$ project anchor boxes into a shared embedding space. The resulting weights, computed with a top-k softmax for stability, encode both spatial proximity and semantic similarity, allowing each tracking query to selectively aggregate information from the most relevant detection queries. Each tracking query then updates its content embedding by aggregating information from the detection queries using the computed attention weights:

$$C_q^{\text{trk}} = \sum_{k \in \mathcal{N}_k(q)} w_{qk} C_k^{\text{det}},$$

where C_k^{det} denotes the content embedding of the k-th detection query, and $\mathcal{N}_k(q)$ denotes the indices of top-k detection queries selected for tracking query q. This update mechanism enables each tracking query to draw identity-relevant cues from the most compatible detections in the current frame, allowing it to adapt to changes in position, appearance, and pose.

Generating Tracking Queries at Training Time. During training, we construct tracking queries directly from ground-truth annotations, following a strat-

egy inspired by DN-DETR [22]. Rather than generating multiple denoised variants per instance, we create a single tracking query for each ground-truth person, which serves as the identity anchor for that individual during training.

To improve robustness to localization errors, we perturb the ground-truth bounding boxes by adding Gaussian noise to their center coordinates (x, y) and size parameters (w, h) before converting them into tracking-query anchors. We further apply stochastic supervision by sampling only a subset of ground-truth instances at each iteration, instead of supplying tracking queries for all annotated instances. As a result, our model becomes robust to imperfect or noisy bounding boxes from previous frames.

User Prompting. At inference time, we optionally allow user-provided box prompts to initialize tracking-query anchors. A user may provide a bounding box (e.g., in the first frame or after a missed track) to initialize or re-initialize a tracking query for the target person, enabling the model to maintain or recover that identity in the remaining frames. DETRAM currently supports bounding-box prompts. Other prompting modalities such as masks, points, or text can be incorporated (when supported) via external mapping or grounding methods.

3.2 DETRAM

Given the encoded frame tokens and the sets of detection and tracking queries, we introduce DETRAM, our unified transformer architecture for end-to-end detection, tracking, and human mesh recovery. Its core contribution is a cross-attention mechanism that jointly reasons over detection, tracking, and prompt queries within a single decoder, enabling identity-consistent multi-person reconstruction, without requiring any external detector or a separate learned tracking module. A schematic of our design is provided in Fig. 3.

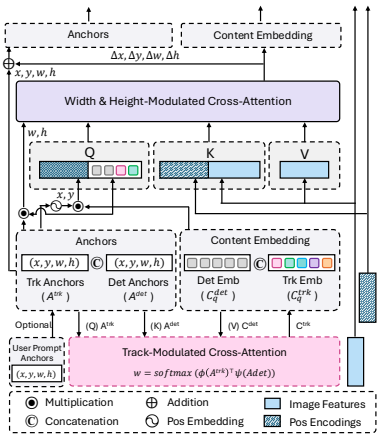

Fig. 3: Our cross-attention design incorporates detection, tracking, and prompt queries. Each query consists of a content embedding and an anchor (learned or provided) defining its positional embedding. TMCA first refreshes tracking/prompt queries from current detections, followed by width- and height-modulated cross-attention over image features.

Shared Decoder for Detection, Tracking, and Prompt Queries. To allow detection and tracking queries to interact seamlessly during decoding, we first concatenate their anchor boxes and content embeddings to form query representations. Detection queries use learned anchors and content embeddings, while tracking queries use propagated anchors with TMCA-refreshed contents and prompt queries are initialized from user-provided anchors. Importantly, this concatenation does not merge them into

a single query representation. Rather, they are decoded in parallel within the same decoder. Positional encodings for the combined queries are then generated by applying an MLP to sinusoidal embeddings of the anchor parameters:

$$A_q = \mathrm{Cat}(A_q^{\mathrm{det}}, A_q^{\mathrm{trk}}), \qquad C_q = \mathrm{Cat}(C_q^{\mathrm{det}}, C_q^{\mathrm{trk}}),$$
$$P_q^{(\cdot)} = \mathrm{MLP}\big(\mathrm{PE}(A_q^{(\cdot)})\big),$$
$$\mathrm{PE}(A) = \mathrm{Cat}\big(\mathrm{PE}(x), \mathrm{PE}(y), \mathrm{PE}(w), \mathrm{PE}(h)\big),$$

where $\mathrm{PE}(\cdot)$ maps a scalar to a $\frac{D}{2}$-dimensional sinusoidal embedding. This formulation allows detection, tracking, and prompt queries to interact within a shared decoder while preserving their distinct supervision. Joint decoding enables tracking queries to attend to detection queries, allowing identity information to flow between existing tracks and newly detected instances within the decoder.

Decoder Architecture. Each transformer decoder layer contains a self-attention module followed by a cross-attention module. While both require queries, keys, and values, they differ in how these components are constructed.

In the *self-attention* module, all queries, keys, and values use the same content embeddings, while positional encodings are added to the queries and keys:

$$Q_q = C_q + P_q, \qquad K_q = C_q + P_q, \qquad V_q = C_q.$$

Conditional Cross-Attention. For cross-attention, we adopt a Conditional-DETR–style formulation [34] that blends semantic and spatial cues to enhance spatial selectivity. This yields spatially conditioned queries and keys:

$$Q_q = \mathrm{Cat}\big(C_q, \mathrm{PE}(x_q, y_q) \cdot \mathrm{MLP}^{(\mathrm{csq})}(C_q)\big),$$
$$K_{x,y} = \mathrm{Cat}\big(F_{x,y}, \mathrm{PE}(x, y)\big),$$
$$V_{x,y} = F_{x,y},$$

where $F_{x,y} \in \mathbb{R}^D$ denotes the encoder feature at spatial location (x, y), and $\mathrm{MLP}^{(\mathrm{csq})}$ modulates positional embeddings based on semantic content. This modulation strengthens the model's ability to correlate query hypotheses with local image evidence, thereby improving both detection precision and identity stability.

3.3 Adding New Humans & Memory Mechanism

We maintain persistent tracking queries with a simple, non-learned track management heuristic on decoder outputs (i.e., not a separate learned tracker) to initialize and update tracks over time. DETRAM also optionally uses a lightweight, parameter-free memory bank as temporal regularization to stabilize identity features over long sequences.

Initializing and Updating Tracks. At the first frame, all detections with confidence above τ_{det} are instantiated as tracking queries and stored in memory. For subsequent frames, a detection is promoted to a new tracking query if

its confidence exceeds τ_{det} and it has low overlap with all existing tracks (IoU $< \tau_{\text{iou}}$). Each tracking query remains active as long as its confidence stays above τ_{det}; if it becomes inactive for more than L_{tol} frames, it is removed from memory. To further manage the life-cycle of tracks, we mark individual tracks as active/inactive/removed based on the following process: (i) tracks with confidence below τ_{det} are marked inactive, and updates to their queries and memory entries are frozen until re-association, (ii) inactive tracks are resumed upon confidence (or association-score) recovery, and (iii) inactive tracks are removed if they remain inactive for more than L_{tol} frames. This mechanism allows the system to dynamically grow or prune the active set of identities, enabling robust handling of entering, exiting, or reappearing individuals.

Key–Value Memory for Long-Term Propagation. To ensure that tracking queries retain identity cues even under occlusion or pose changes, we adopt a simple yet effective key–value memory mechanism inspired by XMem [9] and STCN [10]. Let the memory store keys $m_k \in \mathbb{R}^{B\times(N_t\times M)\times D}$ from previous frames and current query keys $q_k \in \mathbb{R}^{B\times N_t\times D}$. We measure their pairwise similarity by $S = \frac{2\, m_k^\top q_k - \|m_k\|^2}{\sqrt{C_K}}$, where B is the batch size, N_t is the number of active tracks, M is the memory length (number of stored slots), and D is the feature dimension. Each similarity map $S \in \mathbb{R}^{B\times N_t\times M}$ captures how strongly each tracking query attends to previously stored frames. We compute attention weights via a top-k softmax, $\alpha_{b,n,m} = \text{softmax}_m\left(S'_{b,n,m}\right)$, which focuses retrieval on the most relevant memory slots. The final aggregated memory feature is obtained as $\text{mem}_{b,n} = \sum_m \alpha_{b,n,m}\, m_v(b,m)$, where m_v denotes the corresponding memory values. This differentiable retrieval mechanism introduces no additional parameters yet provides temporally stable identity features, enabling DETRAM to maintain consistent tracks through occlusions, rapid motion, and long sequences. The memory bank stores M feature slots per active track, scaling as $O(N_t M D)$ rather than with video length due to the track lifecycle policy.

3.4 Losses

Our training objective integrates detection, reconstruction, and tracking supervision within a unified formulation. Following DETR and DN-DETR [7, 22], we use Hungarian Matching to associate predictions with ground truth (GT) instances using a combination of bounding boxes, projected joints, and confidence scores. We additionally incorporate a denoising mechanism to stabilize learning in the early stages. The total loss is expressed as a weighted sum of three components; the matching loss $\mathcal{L}_{\text{matching}}$, the denoising loss $\mathcal{L}_{\text{dn}}$, and the tracking loss $\mathcal{L}_{\text{trk}}$:

$$\mathcal{L} = \lambda_{\text{matching}}\mathcal{L}_{\text{matching}} + \lambda_{\text{dn}}\mathcal{L}_{\text{dn}} + \lambda_{\text{trk}}\mathcal{L}_{\text{trk}}.$$

Further details regarding the loss terms are provided in the supplementary.

4 Experimental Results

Datasets. Our model is trained on several multi-person datasets, including AGORA [38], BEDLAM [4], COCO [25], and 3DPW [30], and PoseTrack [1,12, 15]. Among these datasets, AGORA, BEDLAM, and 3DPW provide high-quality 3D mesh annotations that are directly used for supervision. For COCO, we rely on pseudo 3D annotations generated by NeuralAnnot [35] and supervise only the projected 2D joints. For PoseTrack, we relabel the data with 3D annotations using the Neural Localizer Fields (NLF) method [43].

Evaluation Protocol. To ensure a fair comparison with existing approaches, we evaluate our framework on PoseTrack [1,15], MuPoTS-3D [31], BEDLAM [4], and 3DPW [30]. Following standard practices [19,39,55], we compute both Mean Vertex Error (MVE) and Mean Per-Joint Position Error (MPJPE), reported before and after applying Procrustes Alignment (PA). All errors, unless otherwise specified, are reported in millimeters (mm).

For the tracking evaluation, we follow the official PoseTrack21 [12] protocol and report standard multi-object tracking metrics, including Multi-Object Tracking Accuracy (MOTA), which combines false positives (FP), missed detections (FN), and ID switches (IDs) into a single accuracy measure. We also compute IDF1, ID Precision (IDP), and ID Recall (IDR) to assess the consistency of identity assignment across frames. We also report HOTA (Higher Order Tracking Accuracy) [29], which jointly measures detection and association accuracy. Higher values indicate more stable temporal associations between predicted tracks and ground-truth trajectories.

Computational Cost and Implementation details. We train our model using eight NVIDIA A100 GPUs with a batch size of 2 per GPU. For evaluating computational efficiency, we measure the average inference time (in seconds) on a single NVIDIA RTX A4000 GPU. We use a perspective camera model, with a fixed fallback camera setting (fixed FOV and centered principal point). We use a two-stage training protocol: first image-based training on AGORA, BEDLAM, COCO, PoseTrack, and 3DPW, then video training with multi-frame clips (e.g. $T = 4$ or 8) on datasets with sequential annotations (e.g. BEDLAM, PoseTrack, and 3DPW). At inference time, the memory bank scales with the number of active tracks, $O(N_t MD)$, rather than video length, requiring only about 12–24 KB per active track (in FP32 format). For video training, we use only temporally contiguous GT track segments and exclude clips with missing/non-continuous GT identities. Additional implementation details can be found in the supplementary.

4.1 Evaluation on PoseTrack21 and MuPoTS-3D

Tracking performance. We evaluate our method on the PoseTrack21 [12] and MuPoTS-3D [31] benchmark to assess multi-person tracking performance. As shown in Tab. 1a, our method achieves the highest overall performance on

Table 1: Evaluation of tracking performance on the PoseTrack21 (left) and MuPoTS-3D (right) benchmarks. *"DETRAM + prompts"* simulates a user-prompting setting where false negatives are specified via box prompts based on ground-truth, leading to improvements across all tracking metrics.

(a) PoseTrack21

Method	MOTA↑	IDF1↑	IDP↑	IDR↑	FP↓	FN↓	FPS↑
With original evaluation method							
TRMOT [50]	47.2	57.3	70.0	46.6	–	–	–
FairMOT [54]	56.3	63.2	81.0	51.8	–	–	–
CorrTrack + ReID [12]	52.0	66.5	72.4	61.4	–	–	–
Tracktor++ [3]	59.5	69.3	76.4	63.5	–	–	–
CoMotion [36]	67.6	77.9	**83.4**	73.0	–	–	–
DETRAM (Ours)	**71.0**	**78.3**	81.6	**74.3**	–	–	–
With ignore region fix from [36]							
4DHumans [14] (reproduced)	56.7	70.9	**87.1**	59.7	**7817**	50652	0.51
CoMotion [36]	71.4	79.5	**87.1**	73.0	8115	30394	5.68
DETRAM (Ours)	**73.6**	**80.6**	85.0	**75.1**	8066	**26620**	**10.87**
DETRAM (Ours) + prompts	**75.7**	**81.8**	85.0	**77.2**	8059	**23369**	**10.87**

(b) MuPoTS-3D

Method	IDs↓	MOTA↑	IDF1↑	HOTA↑
Trackformer [33]	43	24.9	62.7	53.2
Tracktor [3]	53	51.5	70.9	50.3
FlowPose [51]	49	21.4	67.1	41.8
T3DP [40]	38	62.1	79.1	59.2
PHALP [41]	22	66.2	81.4	59.4
ByteTrack [53]	15	73.3	84.6	63.5
TRACE [47]	**0**	86.9	93.4	65.3
CoMotion [36]	**0**	92.1	95.9	69.2
DETRAM (Ours)	**0**	**95.9**	**97.4**	**71.3**
DETRAM (Ours) + prompts	**0**	**96.5**	**98.2**	**71.6**

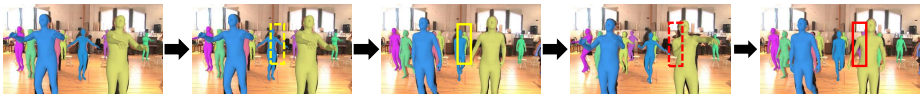

Fig. 4: DETRAM + prompts. User prompts (yellow/red dashed boxes) can be specified when false negatives (missed detections) occur. Propagated boxes (solid line boxes) then follow the person in future frames, recovering all humans.

PoseTrack21, obtaining a MOTA of 73.6 and an IDF1 of 80.6, while also reducing false negatives (FN) compared to previous approaches such as 4DHumans [14] and CoMotion [36]. We also achieve state-of-the-art results on MuPoTS-3D [31] with a MOTA of 95.9, IDF1 of 97.4 and HOTA of 71.3. These improvements indicate better identity preservation and temporal associations.

As DETRAM allows prompting, we simulate a realistic semi-automatic setting to quantify the potential of this capability: whenever a missed detection (false negative) first occurs, we pass in a bounding box derived from the ground-truth annotation for that person (an example is demonstrated in Fig. 4). While this oracle protocol provides an upper bound on the benefit of prompting, the consistent improvements across all tracking metrics (last row of Tab. 1a and Tab. 1b) validate that prompt queries are effectively propagated by the model. Beyond tracking accuracy, DETRAM runs at 10.87 FPS on PoseTrack21 – 1.9× faster than CoMotion [36] and over 21× faster than 4DHumans [14], highlighting the efficiency of our single-stage architecture.

Qualitative results. Fig. 5 presents qualitative comparisons between our method and CoMotion [36] on diverse multi-person scenes. In the first sequence, our model detects more individuals accurately, especially in the background (yellow box). In the second and fourth sequences, our approach maintains stable tracking and consistent reconstruction despite camera movement, frequent occlusions, and limited full-body visibility (red boxes). The third example highlights that our method produces fewer false positives and maintains better identity consistency across frames. Overall, DETRAM demonstrates more reliable multi-person tracking and reconstruction under challenging real-world conditions.

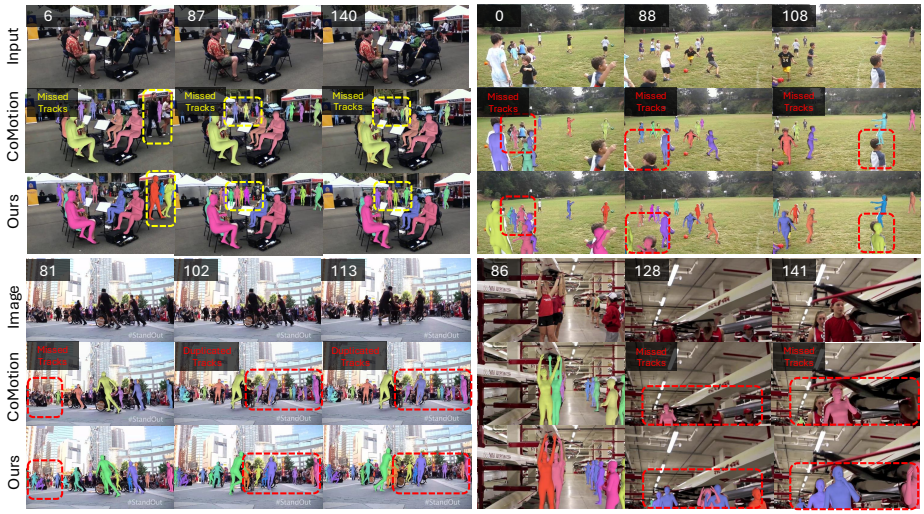

Fig. 5: Qualitative results on the PoseTrack21 dataset show that we are able to detect more humans with fewer false negatives and false positives compared to CoMotion. The numbers in the grey boxes denote the frame index, while the description in grey boxes along with dotted boxes highlight the differences between CoMotion and DETRAM.

Table 2: Evaluation of tracking and reconstruction performance on the BEDLAM dataset. We significantly out-perform the previous methods in key tracking metrics (e.g. IDs and IDF1) and reconstruction metrics (e.g. MPJPE) while attaining comparable inference time and HOTA to MaQ [26] and comparable PA-MPJPE to CoMotion [36].

Method	IDs↓	MOTA↑	IDF1↑	HOTA↑	MPJPE↓	PA-MPJPE↓	PVE↓	Inference Time↓
Fast-RCNN [13] + SMPLer-X [5] + ByteTrack [53]	1660	78.83	81.09	70.41	119.19	52.62	129.22	0.463
Multi-HMR [2] + ByteTrack [53]	515	91.35	92.54	84.59	81.58	42.93	88.88	0.201
MaQ [26]	128	95.15	97.01	**97.73**	79.88	45.56	86.40	**0.027**
CoMotion [36]	885	95.22	97.39	70.94	55.40	**24.50**	62.17	0.239
DETRAM (Ours)	**37**	**99.18**	**99.37**	95.90	**48.68**	31.24	**56.33**	0.069

4.2 Evaluation on BEDLAM

Tracking performance. We evaluate our approach on the BEDLAM [4] benchmark following the evaluation protocol established by MaQ [26]. As shown in Tab. 2, our method achieves the best tracking accuracy, achieving a MOTA of 99.18 and an IDF1 of 99.37, while drastically reducing the number of ID switches to only 37. These results highlight the model's capability to maintain identity consistency and temporal coherence across long and crowded sequences, outperforming all baselines by a clear margin.

Reconstruction performance. In addition to tracking, our method significantly improves 3D reconstruction quality. It achieves lower MPJPE of 48.68 mm, PA-MPJPE of 31.24 mm, and PVE of 56.33 mm compared to previous approaches. This demonstrates the effectiveness of our unified architecture, which jointly optimizes detection, tracking, and 3D reconstruction in a single frame-

work. Although our method achieves superior tracking and reconstruction accuracy overall, the HOTA score is slightly lower than MaQ [26]. We believe this to be due to the differences how 2D bounding boxes are defined, which can reduce HOTA even though the underlying 3D pose and shape estimations are more accurate and consistent across frames. Furthermore, while we out-perform CoMotion [36] on most metrics, it shows a low PA-MPJPE score, which may be due to its larger training datasets (which include DanceTrack and InstaVariety, which we do not train on).

Fig. 6: Qualitative results on BEDLAM, in comparison with MaQ [26]. Note that the MaQ visualizations are taken directly from their paper as they do not provide inference code. DETRAM predicts more accurate human meshes (pose and proportions).

Qualitative results. Fig. 6 shows qualitative comparisons with MaQ [26] on the BEDLAM dataset. Our method demonstrates robust performance even under occlusions, successfully reconstructing multiple interacting people with consistent identities. Compared to MaQ, our approach generates more anatomically plausible human shapes and poses, particularly preserving body proportions. In contrast, MaQ tends to produce overly slim or elongated body shapes. These qualitative observations further validate the advantage of our unified architecture in achieving both accurate 3D pose estimation and stable multi-person tracking.

4.3 Evaluation on 3DPW

Tracking and reconstruction performance. We further evaluate DETRAM on the Dyna3DPW and 3DPW dataset by following the evaluation protocol in [26, 47]. As shown in Tab. 3, DETRAM achieves strong tracking consistency, obtaining the best MOTA (99.8), IDF1 (99.9), and HOTA (79.1), while maintaining zero ID switches. In terms of reconstruction, our method achieves competitive accuracy with an MPJPE of 60.0 mm and PVE of 70.1 mm, outperforming prior methods in overall tracking stability while maintaining 3D reconstruction quality. These results demonstrate that our identity-consistent query formulation

Table 3: Evaluation of tracking and reconstruction performance on Dyna3DPW and 3DPW, respectively, following the protocol in [26, 47].

Method	Dyna3DPW (Tracking)				3DPW (Reconstruction)		
	IDs↓	MOTA↑	IDF1↑	HOTA↑	PA-MPJPE↓	MPJPE↓	PVE↓
BEV [48] + ByteTrack [53]	37	93.6	79.1	59.3	46.9	78.5	92.3
TRACE [47]	1	99.3	99.7	74.7	50.8	80.3	98.1
MaQ [26]	**0**	99.6	99.8	70.8	44.7	72.6	84.9
CoMotion [36]	**0**	95.7	97.8	72.8	**37.3**	60.6	71.2
DETRAM (Ours)	**0**	**99.8**	**99.9**	**79.1**	38.3	**60.0**	**70.1**

Table 4: Ablation study comparing the effects of the tracking query (Track Q) and memory mechanism (Mem Bank) on PoseTrack21 and BEDLAM.

Track Q	Mem Bank	PoseTrack21		BEDLAM			
		MOTA↑	IDF1↑	MOTA↑	IDF1↑	IDs↓	MPJPE↓
✓	✓	**73.6**	**80.6**	**99.18**	**99.37**	37	48.68
✓	-	73.3	80.3	99.10	99.33	**32**	47.33
-	✓	68.9	78.1	98.19	98.90	37	48.27
-	-	68.3	77.1	96.70	97.98	58	**46.91**

Table 5: Stratified analysis of tracking performance on crowded scenes from PoseTrack21.

GT Bin	#Seq	DETRAM (Ours)		CoMotion [36]	
		IDF1 ↑	MOTA ↑	IDF1 ↑	MOTA ↑
1–10	103	87.69	**82.13**	**87.72**	82.05
11–20	44	**79.58**	**74.60**	75.81	69.72
21+	23	**71.66**	**66.83**	70.67	60.53

generalizes well to in-the-wild videos with camera motion. CoMotion achieves a lower PA-MPJPE (37.3 vs. 38.3 mm), likely due to additional training data (DanceTrack and InstaVariety). However, DETRAM outperforms it on MPJPE (60.0 vs. 60.6), PVE (70.1 vs. 71.2), and all tracking metrics.

4.4 Further Analysis

Ablation Study. In Tab. 4 we conduct an ablation study. We show that applying *tracking queries* (Track Q) and the *memory mechanism* (Mem Bank) independently results in modest improvements in tracking performance (MOTA, IDF1), while slightly sacrificing reconstruction performance (MPJPE). Accordingly, we treat the memory mechanism as an optional, parameter-free temporal regularization that stabilizes identity features over long sequences. The memory bank yields a modest gain in MOTA on PoseTrack21 (73.3 to 73.6) with a small runtime cost (11.21 to 10.87 FPS). Its higher BEDLAM MPJPE (47.33 to 48.68) but improved MuPoTS-3D 3DPCK@150 [32] (95.66% to 97.84%) suggests a recall–coverage trade-off. That is, recovering more occluded people can increase the number of difficult-to-reconstruct instances. While its individual effect is modest, combining it with tracking queries yields the best overall tracking performance on both PoseTrack21 and BEDLAM. This is consistent with our design intent: the memory bank is not a primary contribution but a complementary component that provides diminishing returns when tracking queries already maintain strong identity signals. The performance differences are intuitive, as *tracking queries* aid in maintaining identity continuity and can track more individuals than detection queries in crowded scenes, while the *memory mechanism* continues tracking known entities. This is why *tracking queries* are particularly impactful on the PoseTrack21 dataset, which contains numerous

Table 6: Ablation of TMCA and content-conditioned spatial modulation.

Variant	PoseTrack21		BEDLAM			
	MOTA↑	IDF1↑	MOTA↑	IDF1↑	IDs↓	MPJPE↓
w/o Track-Q	68.3	77.1	96.70	97.98	58	46.91
TMCA → direct	71.7	79.6	99.16	99.35	47	49.37
$\text{MLP}^{(\text{csq})}$ → id.	70.2	77.1	98.38	98.05	305	52.73
DETRAM (full)	73.6	80.6	99.18	99.37	37	48.68

entities and entry/exit events. Tracking queries alone produce slightly fewer ID switches on BEDLAM (32 vs. 37), likely due to occasional memory-induced drift from stale embeddings after long occlusions. However, the combined configuration achieves higher overall MOTA and IDF1, indicating that the memory bank's benefits outweigh occasional misassociations.

Crowded Scenes. Challenges in multi-person tracking often arise in crowded scenes. In Tab. 5, we perform a comparison of DETRAM against CoMotion [36] by evaluating tracking performance stratified by the number of people present in a given sequence. We find that in sequences with few people (1–10), DETRAM and CoMotion perform comparably, while DETRAM performs notably better with increasing crowdedness (11–20 and 21+ bins), exhibiting over 10% improvement in MOTA over CoMotion for the most crowded sequences (21+).

TMCA and query modulation ablations. As shown in Tab. 6, replacing TMCA with direct propagation reduces PoseTrack21 MOTA from 73.6 to 71.7, confirming the benefit of refreshing tracking queries from current detections. Replacing $\text{MLP}^{(\text{csq})}$ with an identity mapping increases BEDLAM ID switches from 37 to 305 and worsens MPJPE from 48.68 to 52.73, showing that content-conditioned spatial modulation is critical for stable identity tracking. Overall, TMCA-based tracking trades slightly higher MPJPE for stronger identity consistency.

5 Conclusion

We introduced DETRAM, a unified transformer framework that jointly performs detection, tracking, and human mesh recovery within a single feed-forward pass per frame. By integrating detection, tracking, and prompt queries, it maintains temporal consistency, handles occlusions, and enables controllable tracking. Its framewise design enables online streaming inference with prompt reinitialization. Experiments on PoseTrack21, MuPoTS-3D, 3DPW, and BEDLAM demonstrate improved accuracy and efficiency. Future work will explore explicit temporal smoothness metrics, motion priors for improved temporal coherence, and additional prompting modalities such as text or point-click inputs.

Acknowledgments. This work was supported by NSF grant CRCNS-2309041 and the Harvard Data Science Initiative Trust in Science Fund Award.

References

1. Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., Schiele, B.: Posetrack: A benchmark for human pose estimation and tracking. In: CVPR. pp. 5167–5176 (2018) 10
2. Baradel, F., Armando, M., Galaaoui, S., Brégier, R., Weinzaepfel, P., Rogez, G., Lucas, T.: Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In: ECCV. pp. 202–218. Springer (2024) 2, 3, 4, 12
3. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: ICCV. pp. 941–951 (2019) 11
4. Black, M.J., Patel, P., Tesch, J., Yang, J.: BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: CVPR. pp. 8726–8737 (Jun 2023) 10, 12
5. Cai, Z., Yin, W., Zeng, A., Wei, C., Sun, Q., Yanjun, W., Pang, H.E., Mei, H., Zhang, M., Zhang, L., et al.: Smpler-x: Scaling up expressive human pose and shape estimation. NeurIPS **36**, 11454–11468 (2023) 12
6. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025) 4
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229. Springer (2020) 4, 9
8. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR. pp. 1290–1299 (2022) 4
9. Cheng, H.K., Schwing, A.G.: Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: ECCV. pp. 640–658. Springer (2022) 9
10. Cheng, H.K., Tai, Y.W., Tang, C.K.: Rethinking space-time networks with improved memory coverage for efficient video object segmentation. NeurIPS **34**, 11781–11794 (2021) 9
11. Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., Black, M.J.: Monocular expressive body regression through body-driven attention. In: ECCV (2020) 3
12. Doering, A., Chen, D., Zhang, S., Schiele, B., Gall, J.: Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In: CVPR. pp. 20963–20972 (2022) 10, 11
13. Girshick, R.: Fast r-cnn. In: ICCV. pp. 1440–1448 (2015) 12
14. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4d: Reconstructing and tracking humans with transformers. In: ICCV. pp. 14783–14794 (2023) 2, 3, 11
15. Iqbal, U., Milan, A., Gall, J.: PoseTrack: Joint multi-person pose estimation and tracking. In: CVPR (2017), http://arxiv.org/abs/1611.07727 10
16. Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image. In: CVPR. pp. 5579–5588 (2020) 3
17. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018) 3
18. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: CVPR (2019) 3
19. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: Pare: Part attention regressor for 3d human body estimation. In: ICCV. pp. 11127–11137 (2021) 3, 10

20. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019) 3
21. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: CVPR (2019) 3
22. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: Dn-detr: Accelerate detr training by introducing query denoising. In: CVPR. pp. 13619–13627 (2022) 2, 4, 7, 9
23. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: CVPR (2021) 3
24. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: CVPR. pp. 1954–1963 (2021) 3
25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014) 10
26. Liu, K., Fu, Y., Yuan, W., Lin, J., Li, P., Gu, X., Qiu, L., Wang, H., Dong, Z., Han, X.: Motions as queries: One-stage multi-person holistic human motion capture. In: CVPR. pp. 17529–17539 (2025) 3, 12, 13, 14
27. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. arXiv preprint arXiv:2201.12329 (2022) 2, 4
28. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: a skinned multi-person linear model. ACM Transactions on Graphics (TOG) **34**(6), 1–16 (2015) 4
29. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. International journal of computer vision **129**(2), 548–578 (2021) 10
30. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: ECCV (2018) 10
31. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3d body pose estimation from monocular rgb input. In: 3DV (2018) 10, 11
32. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3d pose estimation from monocular rgb. In: 2018 international conference on 3D vision (3DV). pp. 120–130. IEEE (2018) 14
33. Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. In: CVPR. pp. 8844–8854 (2022) 11
34. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J.: Conditional detr for fast training convergence. In: ICCV. pp. 3651–3660 (2021) 8
35. Moon, G., Choi, H., Lee, K.M.: Neuralannot: Neural annotator for 3d human mesh training sets. In: CVPR. pp. 2299–2307 (2022) 10
36. Newell, A., Hu, P., Lipson, L., Richter, S., Koltun, V.: Comotion: Concurrent multi-person 3d motion. In: ICLR (2025) 3, 11, 12, 13, 14, 15
37. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 5
38. Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J.: Agora: Avatars in geography optimized for regression analysis. In: CVPR. pp. 13468–13478 (2021) 10

39. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR. pp. 10975–10985 (2019) 10
40. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Malik, J.: Tracking people with 3d representations. NeurIPS **34**, 23703–23713 (2021) 2, 3, 11
41. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Malik, J.: Tracking people by predicting 3d appearance, location and pose. In: CVPR. pp. 2740–2749 (2022) 2, 3, 11
42. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net++: Multi-person 2d and 3d pose detection in natural images. TPAMI (2019) 3
43. Sárándi, I., Pons-Moll, G.: Neural localizer fields for continuous 3d human pose and shape estimation. In: NeurIPS (2024) 10
44. Su, C., Ma, X., Su, J., Wang, Y.: Sat-hmr: Real-time multi-person 3d mesh estimation via scale-adaptive tokens. In: CVPR. pp. 16796–16806 (2025) 2, 3, 4
45. Sun, Q., Wang, Y., Zeng, A., Yin, W., Wei, C., Wang, W., Mei, H., Leung, C.S., Liu, Z., Yang, L., et al.: Aios: All-in-one-stage expressive human pose and shape estimation. In: CVPR. pp. 1834–1843 (2024) 2, 3, 4
46. Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, one-stage, regression of multiple 3d people. In: ICCV (2021) 3
47. Sun, Y., Bao, Q., Liu, W., Mei, T., Black, M.J.: Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In: CVPR (2023) 11, 13, 14
48. Sun, Y., Liu, W., Bao, Q., Fu, Y., Mei, T., Black, M.J.: Putting people in their place: Monocular regression of 3d people in depth. In: CVPR. pp. 13243–13252 (2022) 3, 14
49. Wang, Y., Sun, Y., Patel, P., Daniilidis, K., Black, M.J., Kocabas, M.: Prompthmr: Promptable human mesh recovery. In: CVPR. pp. 1148–1159 (2025) 3
50. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: ECCV. pp. 107–122. Springer (2020) 11
51. Xiu, Y., Li, J., Wang, H., Fang, Y., Lu, C.: Pose Flow: Efficient online pose tracking. In: BMVC (2018) 11
52. Zanfir, A., Marinoiu, E., Sminchisescu, C.: Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In: CVPR. pp. 2148–2157 (2018) 3
53. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: ECCV (2022) 11, 12, 14
54. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. IJCV **129**(11), 3069–3087 (2021) 11
55. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: ICCV. pp. 11656–11665 (2021) 10