AnthroTAP: Learning Point Tracking with Real-World Motion – Supplementary Materials –

Inès Hyeonsu Kim[1,3*] Seokju Cho[1*] Jahyeok Koo[1] Junghyun Park[1] Jiahui Huang[2]
Honglak Lee[3,4] Joon-Young Lee[2] Seungryong Kim[1]

[1]KAIST AI [2]Adobe Research [3]University of Michigan [4]LG AI Research

A. Additional ablations and analyses

Table 1: Ablation on training video length.

	Video Length	DAVIS AJ	DAVIS $< \delta^x_{avg}$	DAVIS OA
(I)	24	<u>64.7</u>	76.8	**89.4**
(II)	48	**64.8**	**77.3**	<u>89.1</u>
(III)	64	64.5	<u>77.1</u>	88.9

Ablation on training video length. In Table 1, we investigate the influence of training video length on model performance. For this analysis, we fine-tuned the LocoTrack [3] base model using our annotated videos, experimenting with distinct clip lengths. The results demonstrate that utilizing video clips of 48 frames yields the best performance in terms of both AJ and $< \delta^x_{avg}$.

Table 2: Ablation on frame dilation.

	Frame Dilation	DAVIS AJ	DAVIS $< \delta^x_{avg}$	DAVIS OA
(I)	1	<u>64.6</u>	76.7	**89.6**
(II)	2	**64.8**	**77.3**	<u>89.1</u>
(III)	3	64.2	<u>76.9</u>	89.0

Ablation on frame dilation. During training, we adjust the frame rate to control the motion speed by sampling video frames with different dilation factors. In Table 2, we test dilations of 1, 2, and 3, corresponding to 1×, 2×, and 3× faster motion, respectively. When the dilation is too large, the gap between adjacent frames becomes too wide. We found that a dilation factor of 2 yields the best performance in both AJ and $< \delta^x_{avg}$.

Ablation on training data scale. In Table 3, we examine how the amount of additional training data influences model performance. Starting from our full dataset containing 1,400 videos, we evaluate reduced subsets of 500 and 250 videos. The results show that even when using only a fraction of the full dataset, our additional videos consistently improve performance over the LocoTrack baseline across all metrics. This demonstrates that our dataset provides strong supervision signals, and meaningful gains can be achieved even with limited data.

Table 3: Ablation on using different amount of extra data.

Method	DAVIS AJ	DAVIS $< \delta^x_{avg}$	DAVIS OA
LocoTrack (384 × 512)	64.8	77.4	86.2
LocoTrack + **Ours** (250 videos)	65.6	78.5	**87.3**
LocoTrack + **Ours** (500 videos)	<u>65.7</u>	<u>78.7</u>	<u>87.2</u>
LocoTrack + **Ours** (full)	**65.9**	**78.9**	**87.3**

B. Additional visualization

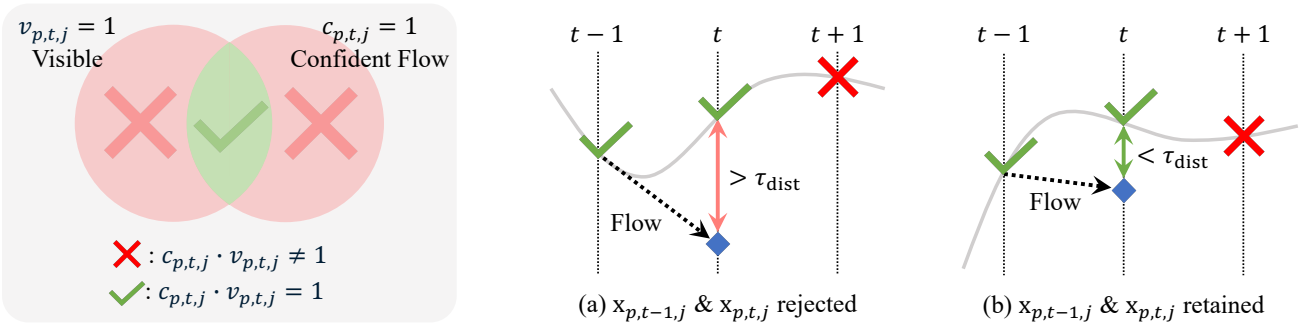
We visualize videos annotated using our pipeline in Figure 3 and Figure 4. The visualizations highlight the complex motion present in the dataset.

C. Filtering outlier trajectories with optical flow

While HMR methods [1, 4, 5] have become increasingly robust to challenges such as motion blur, rapid movements, and varying lighting conditions due to their strong human priors, they may still produce occasional errors or inconsistencies in mesh predictions under extreme conditions. Furthermore, our ray-casting-based visibility prediction is limited to occlusions caused by other modeled humans and does not account for occlusions from general scene objects. In addition, the parametric nature of the SMPL model makes it less effective at capturing the motion of highly deformable or loosely attached clothing and accessories.

To mitigate the impact of these potential inaccuracies on our pseudo-labels and thus improve the quality of training data, we introduce an optical flow-based filtering stage, illustrated in Figure 1. Optical flow [6, 9, 10] is particularly well-suited for this task due to its established accuracy in

*Equal contribution.



(a) $x_{p,t-1,j}$ & $x_{p,t,j}$ rejected (b) $x_{p,t-1,j}$ & $x_{p,t,j}$ retained

Figure 1: Filtering erroneous tracks with optical flow. We filter trajectories predicted from the human mesh using optical flow. First, we retain points that are both considered visible with ray-casting and have confident optical flow at the predicted position ($c_{p,t,j} \cdot v_{p,t,j} = 1$), denoted as $\checkmark$. If the difference between the trajectory predicted from the human mesh and the optical flow exceeds a threshold τ_{dist}, the point is considered erroneous, as depicted in (a), otherwise retained as in (b). We omit the normalization process for clarity.

estimating dense motion between adjacent frames. Over very short temporal windows, these flow-based predictions can provide strong local motion cues to validate or identify discrepancies in the HMR-derived trajectories, especially when dealing with complex non-rigid deformations or potential tracking drift [2, 8].

The primary objective of this stage is not to replace the HMR-derived tracks but to identify and remove potentially erroneous segments or entire trajectories. Our goal is to achieve a high true positive rate for valid track segments while maintaining a sufficiently high true negative rate for incorrect ones. We operate under the assumption that training a point tracker with a smaller set of highly accurate pseudo-labels is more beneficial than using a larger set contaminated with significant errors.

Identify confident optical flow. For each pair of consecutive frames (I_t, I_{t+1}), we compute both the forward optical flow map from I_t to I_{t+1}, denoted $\mathbf{F}_{t\to t+1}$, and the backward optical flow map $\mathbf{F}_{t+1\to t}$. We use $\mathbf{f}(\mathbf{F}, \mathbf{x})$ to denote the process of sampling the flow vector from map $\mathbf{F}$ at subpixel location $\mathbf{x}$ using bilinear interpolation. For notational simplicity, we define:

$$\boldsymbol{\xi}_{p,t,j} = \mathbf{f}(\mathbf{F}_{t\to t+1}, \mathbf{x}_{p,t,j}), \tag{1}$$

$$\boldsymbol{\zeta}_{p,t,j} = \mathbf{f}(\mathbf{F}_{t+1\to t}, \mathbf{x}_{p,t,j} + \boldsymbol{\xi}_{p,t,j}). \tag{2}$$

To assess the reliability of the optical flow estimation itself at a given point $\mathbf{x}_{p,t,j}$, we apply a forward-backward consistency check [7]. Let $c_{p,t,j}$ denote the binary indicator of optical flow reliability at point $\mathbf{x}_{p,t,j}$ in frame I_t. A point $\mathbf{x}_{p,t,j}$ is considered to have reliable flow ($c_{p,t,j} = 1$) if the L_2 distance between the original point location and the location obtained after warping to frame $t+1$ and back to frame t is below a predefined threshold δ_{cons}:

$$c_{p,t,j} = \mathbb{1}\left[\|\boldsymbol{\xi}_{p,t,j} + \boldsymbol{\zeta}_{p,t,j}\|_2 < \delta_{\text{cons}}\right]. \tag{3}$$

Find erroneous trajectories. For each point $\mathbf{x}_{p,t,j}$ within a pseudo-labeled trajectory $\mathcal{X}_{p,j}$ (where t and $t+1$ are both in $\mathcal{T}_{p,j}$), we examine the predicted motion to the next frame, $\mathbf{x}_{p,t+1,j}$. The displacement vector derived from the HMR pseudo-label is $\Delta\mathbf{x}^{\text{HMR}}_{p,t,j} = \mathbf{x}_{p,t+1,j} - \mathbf{x}_{p,t,j}$. We compare this to the optical flow displacement $\boldsymbol{\xi}_{p,t,j}$ (previously defined as the forward flow at location $\mathbf{x}_{p,t,j}$).

To compare these two displacement vectors, $\Delta\mathbf{x}^{\text{HMR}}_{p,t,j}$ and $\boldsymbol{\xi}_{p,t,j}$, robustly, especially in cases of large motion, we normalize them. Let $L_{\text{short}} = \min(\|\Delta\mathbf{x}^{\text{HMR}}_{p,t,j}\|_2, \|\boldsymbol{\xi}_{p,t,j}\|_2) + \epsilon_{\text{norm}}$, where ϵ_{norm} is a small positive constant added to prevent division by zero. The normalized vectors are:

$$\Delta\widehat{\mathbf{x}}^{\text{HMR}}_{p,t,j} = \frac{\Delta\mathbf{x}^{\text{HMR}}_{p,t,j}}{L_{\text{short}}}, \quad \widehat{\boldsymbol{\xi}}_{p,t,j} = \frac{\boldsymbol{\xi}_{p,t,j}}{L_{\text{short}}}. \tag{4}$$

A point transition from frame t to $t+1$ is flagged as potentially erroneous if the L_2 distance between these normalized displacement vectors exceeds a threshold τ_{dist}. We define an indicator variable $e_{p,t,j}$ for this:

$$e_{p,t,j} = \mathbb{1}\left[\|\Delta\widehat{\mathbf{x}}^{\text{HMR}}_{p,t,j} - \widehat{\boldsymbol{\xi}}_{p,t,j}\|_2 > \tau_{\text{dist}}\right]. \tag{5}$$

Trajectory rejection. We observed that query points located on regions not well captured by the SMPL model, such as loose clothing or hair, often result in a large number of transitions being flagged as erroneous. Nonetheless, even in these cases, certain frames may produce transitions that closely follow the predicted motion and are not flagged as erroneous. To robustly filter such trajectories, we evaluate the proportion of transitions that are flagged as erroneous. For a given trajectory $\mathcal{X}_{p,j}$ (defined by the ordered sequence of frame indices $\mathcal{T}_{p,j}$), we calculate the fraction of its transitions that are flagged as erroneous. This ratio, $R_{p,j}$, is computed by summing over all transitions from a frame t_k to its successor t_{k+1} within the sequence $\mathcal{T}_{p,j}$. Only transitions where the point at the starting frame t_k is estimated as visible ($v_{p,t_k,j} = 1$) by ray casting and the optical flow is deemed reliable ($c_{p,t_k,j} = 1$) are included in this calcula-

tion:

$$R_{p,j} = \frac{\sum_{k=1}^{|\mathcal{T}_{p,j}|-1} e_{p,t_k,j} \cdot v_{p,t_k,j} \cdot c_{p,t_k,j}}{\sum_{k=1}^{|\mathcal{T}_{p,j}|-1} v_{p,t_k,j} \cdot c_{p,t_k,j} + \epsilon_{\text{ratio}}}, \quad (6)$$

where (t_k, t_{k+1}) denotes the k-th pair of consecutive frame indices in the ordered set $\mathcal{T}_{p,j}$. The term $e_{p,t_k,j}$ indicates that the transition starting at frame t_k (and ending at t_{k+1}) is erroneous. ϵ_{ratio} is a small constant to ensure numerical stability in case the denominator is zero. A trajectory $\mathcal{X}_{p,j}$ is ultimately rejected if this erroneous transition ratio $R_{p,j}$ exceeds a predefined threshold τ_{ratio}.

For trajectories that pass this trajectory-level filtering, any individual point-pair transition $(\mathbf{x}_{p,t,j}, \mathbf{x}_{p,t+1,j})$ still flagged as erroneous ($e_{p,t,j} = 1$) is considered unreliable. For the purpose of training a point tracking model, such individual erroneous transitions are excluded.

D. Trajectory complexity and diversity analysis

To quantitatively evaluate the characteristics of generated pseudo-labeled trajectories $\mathcal{X}_{p,j}$, we employ metrics for trajectory complexity and dataset diversity. These metrics help in understanding the nature of the motion patterns captured. All trajectories are sequences of 2D points $\mathbf{x}_{p,t,j}$ over time t. Calculations are performed on contiguous visible segments of these trajectories, and points are assumed to be normalized by frame dimensions.

Trajectory complexity. Trajectory complexity is quantified using the mean angular acceleration magnitude. This metric captures the rate of change in the direction of motion, highlighting non-linear movements and directional variations. For a given visible segment of a trajectory $\mathcal{X}_{p,j}$, consisting of an ordered sequence of points $(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_K)$ (where $\mathbf{x}_k$ corresponds to a point $\mathbf{x}_{p,t_k,j}$ from the trajectory, requiring at least 4 points with $K \geq 3$ for at least one angular acceleration value), and assuming a constant time step Δt between frames:

First, a sequence of velocity vectors $\mathbf{u}_k$ between consecutive points is computed:

$$\mathbf{u}_k = \mathbf{x}_{k+1} - \mathbf{x}_k \quad \text{for } k = 0, \ldots, K-1. \quad (7)$$

The signed turning angle θ_k at point $\mathbf{x}_{k+1}$ (i.e., between vectors $\mathbf{u}_k$ and $\mathbf{u}_{k+1}$) is calculated as:

$$\theta_k = \text{atan2}(\mathbf{u}_k^{(1)}\mathbf{u}_{k+1}^{(2)} - \mathbf{u}_k^{(2)}\mathbf{u}_{k+1}^{(1)}, \mathbf{u}_k \cdot \mathbf{u}_{k+1}), \quad (8)$$

for $k = 0, \ldots, K-2$. This sequence of angles is then unwrapped to obtain a continuous sequence of angles $\Phi = (\phi_0, \phi_1, \ldots, \phi_{K-2})$ by adding multiples of $\pm 2\pi$ to eliminate jumps. The sequence of angular velocities ω_k is computed from the unwrapped turning angles:

$$\omega_k = \frac{\phi_k}{\Delta t} \quad \text{for } k = 0, \ldots, K-2. \quad (9)$$

The angular accelerations α_k are then found by taking the difference between consecutive angular velocities:

$$\alpha_k = \frac{\omega_{k+1} - \omega_k}{\Delta t} \quad \text{for } k = 0, \ldots, K-3. \quad (10)$$

The complexity for the segment, C_{seg}, is the mean of the magnitudes of these angular accelerations:

$$C_{\text{seg}} = \frac{1}{K-2} \sum_{k=0}^{K-3} |\alpha_k|. \quad (11)$$

The complexity for an entire trajectory $\mathcal{X}_{p,j}$, denoted $C_{\mathcal{X}_{p,j}}$, is the average of C_{seg} over all its valid visible segments. The overall dataset complexity is the mean of $C_{\mathcal{X}_{p,j}}$ across all trajectories.

Trajectory diversity. Trajectory diversity is assessed by computing the mean standard deviation of centered trajectories from a mean centered trajectory. This measures the spatial variability of trajectories in terms of their shape and relative motion, independent of their absolute starting positions.

Consider a set of N trajectories $\{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_N\}$ within a single video, where $\mathcal{X}_i$ refers to a specific $\mathcal{X}_{p,j}$, and $\mathcal{X}_i(t)$ denotes the 2D coordinate $\mathbf{x}_{p,t,j}$ for the i-th trajectory at frame t. First, each trajectory $\mathcal{X}_i(t)$ is centered by subtracting the coordinates of its first visible point $\mathcal{X}_i(t_{i,\text{first\_vis}})$ from all its subsequent visible points. Let this centered trajectory be $\mathcal{X}'_i(t)$:

$$\mathcal{X}'_i(t) = \mathcal{X}_i(t) - \mathcal{X}_i(t_{i,\text{first\_vis}}) \quad \text{for visible } t \geq t_{i,\text{first\_vis}}. \quad (12)$$

Points where the trajectory is occluded or prior to $t_{i,\text{first\_vis}}$ are considered invalid. A mean centered trajectory $M'(t)$ is computed by averaging the coordinates of all valid centered trajectories at each frame t:

$$M'(t) = \frac{1}{|\mathcal{V}_t|} \sum_{i \in \mathcal{V}_t} \mathcal{X}'_i(t), \quad (13)$$

where $\mathcal{V}_t$ is the set of indices of trajectories having valid centered data at frame t. For each centered trajectory $\mathcal{X}'_i$, the mean squared Euclidean distance (MSD_i) from $M'(t)$ is calculated over all frames $t \in \mathcal{T}_i$ where both $\mathcal{X}'_i(t)$ and $M'(t)$ are valid:

$$MSD_i = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \|\mathcal{X}'_i(t) - M'(t)\|_2^2. \quad (14)$$

The standard deviation for trajectory i is then $SD_i = \sqrt{MSD_i}$. The diversity score D_{video} for the video (or dataset) is the mean of these individual trajectory standard deviations:

$$D_{\text{video}} = \text{mean}(SD_1, SD_2, \ldots, SD_N). \quad (15)$$

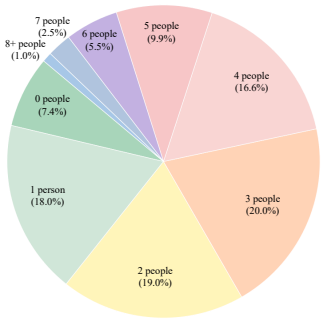


Figure 2: The number of people per frame. Anthro-LD, the dataset generated with our pipeline, contains videos where most scenes include more than one person, contributing to the complexity of the trajectories.

E. Additional data statistics

Statistics on the number of people in training videos. In Figure 2, we investigate the number of people per frame in the generated dataset. Most frames contain more than one person (82%), which can lead to complex trajectories due to occlusions between individuals.

F. Limitations and social impact

While our pseudo-labels provide substantial performance gains when used to train point tracking models, the filtering pipeline is designed conservatively to remove erroneous trajectories, which means some potentially valid trajectories may be rejected. Furthermore, visible points in the video could be mistakenly identified as occluded. This inaccuracy regarding occlusion status led us to decide against supervising occlusion directly, and instead focus only on position. Although a dataset with inaccurate occlusion information can still be helpful for training, its utility as a benchmark is limited.

Social impact. This work significantly enhances point tracking accessibility and efficiency by automating data generation, thereby reducing reliance on costly manual annotation and extensive computational resources. This advancement can benefit robotics, 3D/4D reconstruction, and video editing. Open-sourcing the dataset and pipeline will promote collaborative research on point tracking. While not explicitly discussed, potential misuse for surveillance should be considered.

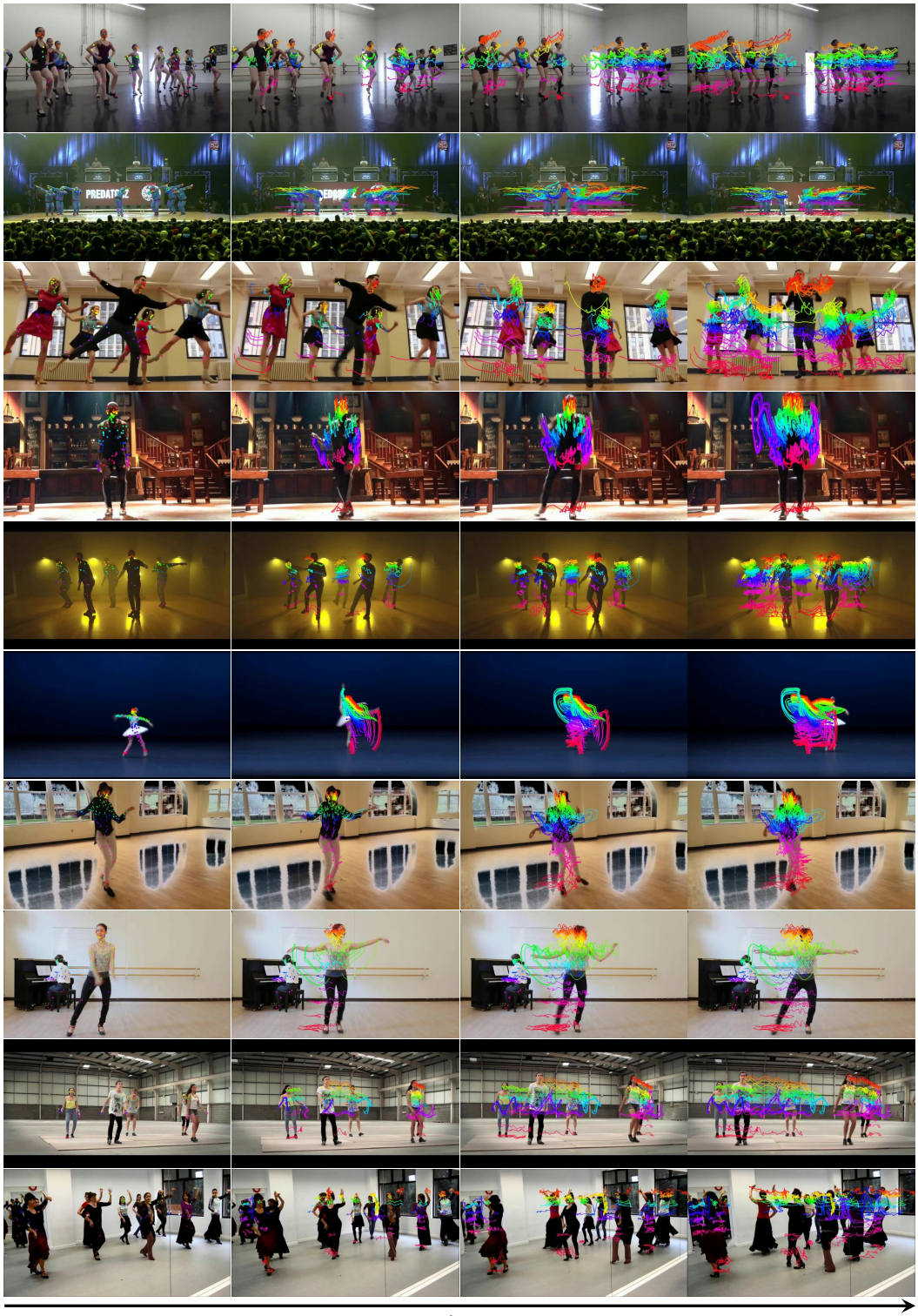


Figure 3: Additional visualization of videos annotated with our pipeline. In this example, we visualize trajectories extracted using our pipeline.

Figure 4: Additional visualization of videos annotated with our pipeline. In this example, we visualize trajectories extracted using our pipeline.

References

[1] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 1

[2] Seokju Cho, Jiahui Huang, Seungryong Kim, and Joon-Young Lee. Flowtrack: Revisiting optical flow for long-range dense tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19268–19277, 2024. 2

[3] Seokju Cho, Jiahui Huang, Jisu Nam, Honggyu An, Seungryong Kim, and Joon-Young Lee. Local all-pair correspondence for point tracking. In *European Conference on Computer Vision*, pages 306–325. Springer, 2024. 1

[4] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1323–1333, 2024. 1

[5] 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. 1

[6] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *European conference on computer vision*, pages 668–685. Springer, 2022. 1

[7] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[8] Michal Neoral, Jonáš Šerých, and Jiří Matas. Mft: Long-term tracking of every pixel. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6837–6847, 2024. 2

[9] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 1

[10] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*, pages 36–54. Springer, 2024. 1