Robust Point Tracking with Epipolar Constraints

Adithya Narayan[*,1] Tanisha Gupta[*,1]
Lamia Alsalloom[*,1]
[1]Carnegie Mellon University
{anaraya2, tanishag, lalsallo}@andrew.cmu.edu

Abstract

We present a geometry-driven, robust point-tracking framework that improves tracking accuracy in videos by enforcing epipolar constraints. While modern trackers such as CoTracker achieve impressive results, they often exhibit geometric drift and produce correspondences that violate fundamental principles of multi-view geometry. We address this in two complementary ways: (1) a sequential refinement module that corrects correspondences frame-by-frame to satisfy epipolar constraints, and (2) a CoTracker finetuning strategy that incorporates a soft epipolar loss on rigid regions of the scene. Both our post-processing approach and our finetuned model reduce geometric drift and improve multi-view consistency, particularly on long sequences with significant camera motion. Although enforcing geometric consistency can slightly reduce standard tracking metrics, it yields significantly lower geometric error on long-horizon videos. Importantly, our post-processing framework is model-agnostic and can be integrated with any point-tracking system.

1. Introduction

Point tracking in videos is a fundamental problem in computer vision with applications ranging from motion analysis and video editing to 3D reconstruction and augmented reality. Recent advances in deep learning have led to highly accurate tracking systems such as CoTracker [10], which can track dense point correspondences across long video sequences. However, despite their impressive performance, these learned trackers can still produce tracks that violate fundamental geometric constraints, particularly in scenarios with significant camera motion.

Geometric Drift in Point Tracking When tracking points across multiple frames, correspondences must satisfy the epipolar constraint: corresponding points in two views lie on corresponding epipolar lines defined by the fundamental matrix. While deep learning trackers learn powerful appearance-based features, they may drift geometrically over time, accumulating errors that violate these constraints. This geometric drift becomes particularly problematic in long sequences or when tracking through challenging conditions such as occlusions, motion blur, or textureless regions. We see a quantitative example of this in Figure 2. We also believe that in general, epipolar constraints offer a smaller search space for models to find correspondences in. We believe this also helps the model reduce drift.

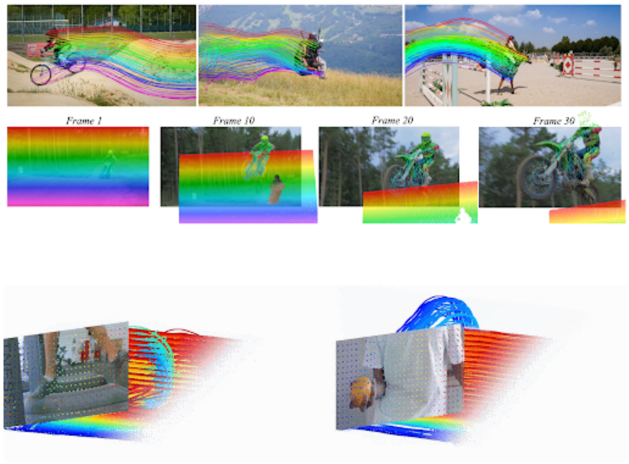


Figure 1. Point tracking can cover both foreground and background points and cover both rigid and non-rigid transformations.

Epipolar Constraints for Refinement The fundamental matrix encodes the geometric relationship between two views of a scene. This gives us a constraint for validating and correcting point correspondences. Given a fundamental matrix $\mathbf{F}$ between two frames, any correspondence (x_0, x_t) must satisfy $x_t^T \mathbf{F} x_0 = 0$, meaning point x_t lies on the epipolar line $\mathbf{F}x_0$. In our iterative approach, we leverage this constraint to refine tracks by iteratively correcting correspondences to minimize their distance to epipolar lines.

[*] Equal contribution, alphabetical by last name.

The contributions of our work are summarized as follows:

- We present an iterative epipolar refinement framework that post-processes tracks from any point tracker to enforce geometric consistency. The method uses RANSAC-based fundamental matrix estimation and distance-based correction to iteratively refine correspondences.
- We introduce a teacher-student based soft-epipolar constraint finetuning that jointly minimizes epipolar errors across all frame pairs while teaching the student model to implicitly seperate static and dynamic components in the scene.

2. Related Work

Point Tracking Methods Point tracking has been extensively studied in computer vision, with approaches ranging from classical optical flow methods [8] to modern deep learning-based trackers. Recent transformer-based models like CoTracker [10], TAPIR [3], and PIPs [6] have achieved state-of-the-art performance by jointly tracking multiple points and leveraging temporal context. However, these methods primarily rely on appearance matching and learned features, without explicitly enforcing geometric constraints that govern multi-view correspondences on static components of the scene.

Epipolar Geometry and Fundamental Matrix The fundamental matrix is a fundamental concept in multi-view geometry that encodes the relationship between two views of a scene [7]. It relates corresponding points through the epipolar constraint: $x'^T\mathbf{F}x = 0$, where x and x' are corresponding points in homogeneous coordinates. The fundamental matrix can be estimated from point correspondences using methods such as the 8-point algorithm [11] or robust estimation techniques like RANSAC [4]. Our work leverages these classical techniques to refine modern deep learning trackers.

Geometric Consistency in Tracking Several works have explored incorporating geometric constraints into tracking pipelines. Some methods use epipolar constraints for outlier rejection [13] or as part of structure-from-motion pipelines [12]. However, most deep learning trackers treat tracking as a purely appearance-based problem. Recent work has explored combining learned features with geometric constraints [14], but these use heavy 3D triangulation based strategies to implicitly model the scene geometry.

Post-Processing and Refinement Post-processing techniques have been widely used to improve tracking results, including temporal smoothing [2], Kalman filtering [9], and optical flow refinement [1]. However, these methods typically focus on temporal consistency rather than geometric constraints. Our method specifically addresses geometric drift by enforcing epipolar constraints, which is particularly important for applications requiring geometric accuracy such as 3D reconstruction or camera pose estimation.

3. Methodology

3.1. Post-Processing Approach

As shown in Figure 3, our epipolar refinement framework consists of two main stages that are iterated repeatedly: estimation of the fundamental matrix per frame pair and outlier correction. The pipeline proceeds through the following steps. First, we estimate fundamental matrices between frame 0 and each subsequent frame using RANSAC. Next, we iteratively refine tracks by correcting correspondences to minimize their distance to epipolar lines. Then, we restimate F_{ij} with a tighter threshold until convergence.

As for our second approach, we take inspiration from the soft epipolar constraint defined in ROMO (Robust Motion Segmentation Improves Structure from Motion) [5]. We use this to define L_{epi} that we combine with $L_{cotrack}$ (or $L_{dynamic}$)to finetune our model on the Tap-Vid DAVIS dataset.

3.1.1. Fundamental Matrix Estimation

Given tracks from a point tracker, we first estimate fundamental matrices between frame 0 and each subsequent frame $t \in \{1, 2, \ldots, T-1\}$. For each frame pair $(0, t)$, we extract mutually visible correspondences and use RANSAC to robustly estimate the fundamental matrix.

The RANSAC algorithm proceeds as follows. We randomly sample 8 point correspondences that are **mutually visible in both frames**and compute the fundamental matrix using the 8-point algorithm. We count inliers, which are correspondences with epipolar error below threshold τ. This process repeats until the confidence threshold is reached or maximum iterations are exceeded. Finally, we refine the fundamental matrix using all inliers.

The epipolar error for a correspondence (x_0, x_t) is computed as:

$$d_{\text{epipolar}} = \frac{|x_t^T \mathbf{F} x_0|}{\sqrt{(\mathbf{F}x_0)_1^2 + (\mathbf{F}x_0)_2^2}} \tag{1}$$

where $(\mathbf{F}x_0)_i$ denotes the i-th component of the epipolar line vector.

3.1.2. Sequential Refinement

After estimating the fundamental matrices, we iteratively refine tracks to satisfy epipolar constraints. For each iteration $k \in \{1, 2, \ldots, K\}$, we re-estimate fundamental matrices from the current tracks using RANSAC. For each frame pair $(0, t)$ and each visible correspondence n, we compute the epipolar line in frame t as $\mathbf{l}' = \mathbf{F}\mathbf{x}_0^{(n)}$.

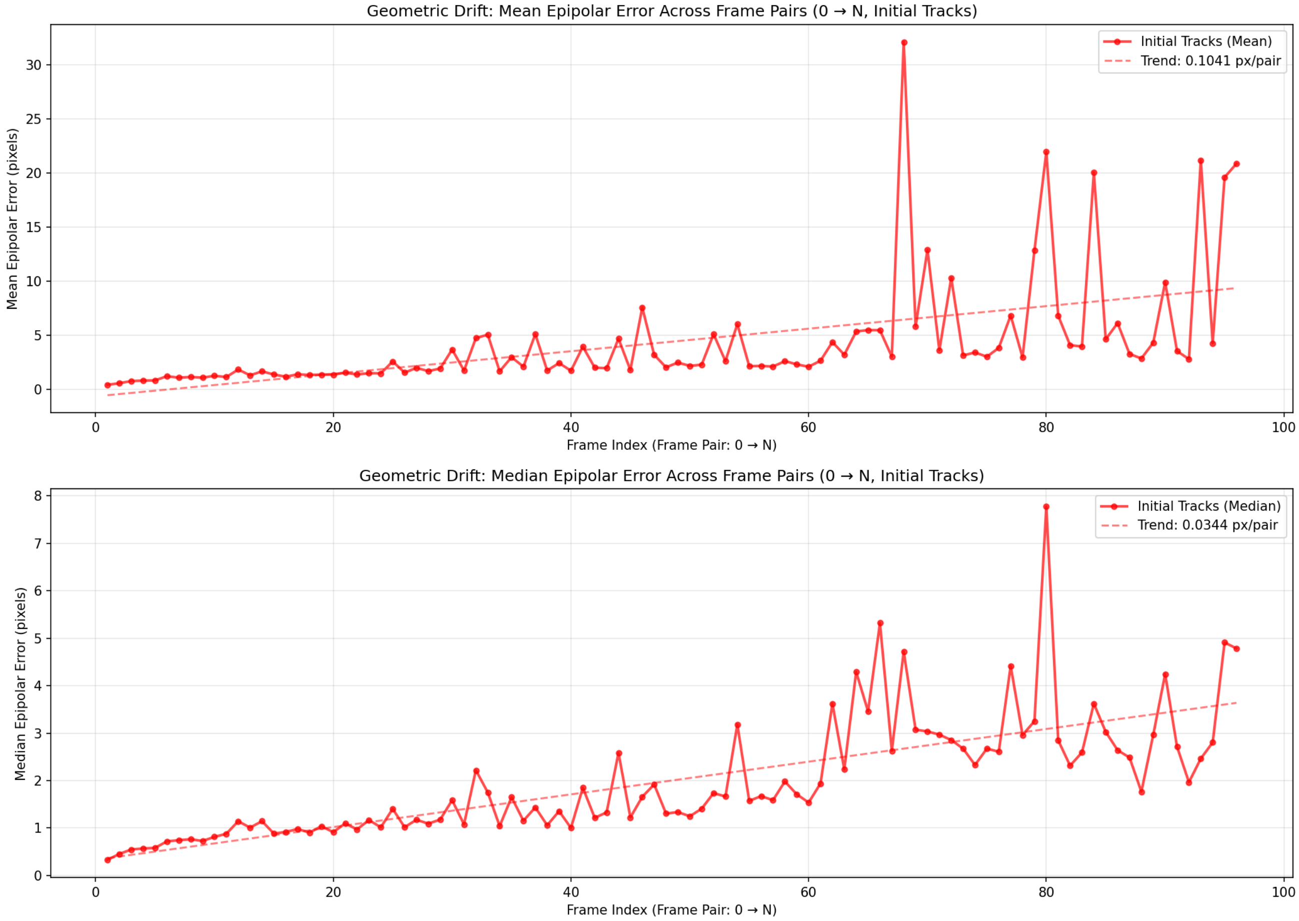


Figure 2. A quantitative look at epipolar drift over time. This is averaged over 300 samples across around 100 frames. Notice that the mean (top) and the median (bottom) increase over time. This suggests that there's a consistent increase in geometric drift over time.

Rather than directly projecting $\mathbf{x}_t^{(n)}$ onto this line, we now perform a **SIFT-based local search**: we extract a SIFT descriptor at the reference location in frame 0 and search for the best-matching SIFT feature within a **20-pixel window** centered on the epipolar line in frame t. The resulting matched location is taken as the refined correspondence $\mathbf{x}_t^{(n)*}$.

If no reliable SIFT match is found, we fall back to geometric refinement by computing the projection of $\mathbf{x}_t^{(n)}$ onto the epipolar line. Given $\mathbf{l}' = [a, b, c]^T$ (where $ax+by+c = 0$) and point (x_0, y_0), the closest point on the line is:

$$x_{\text{proj}} = \frac{b(bx_0 - ay_0) - ac}{a^2 + b^2} \tag{2}$$

$$y_{\text{proj}} = \frac{a(ay_0 - bx_0) - bc}{a^2 + b^2}. \tag{3}$$

We then **directly update the correspondence** as

$$\mathbf{x}_t^{(n)} \leftarrow \mathbf{x}_t^{(n)*}.$$

3.2. Weak Supervision Approach

Our finetuning pipeline extends the standard CoTracker training procedure to incorporate weak supervision signals derived from epipolar-informed motion masks computed using ROMO. Instead of explicitly segmenting the input using the static and dynamic motion masks we get from ROMO, we employ a teacher-student setup: the student learns to separate static and dynamic components while correcting geometric drift in static points based on the teacher's predictions.

Inspired by ROMO, given two consecutive frames I_t and I_{t+1}, we compute forward and backward optical flow, $f_{t\to t+1}$ and $f_{t+1\to t}$, to establish coarse correspondences:

$$\mathbf{x}_{t+1}^{\text{flow}} \approx \mathbf{x}_t + f_{t\to t+1}(\mathbf{x}_t), \quad \mathbf{x}_t^{\text{flow}} \approx \mathbf{x}_{t+1} + f_{t+1\to t}(\mathbf{x}_{t+1}). \tag{4}$$

Forward-backward consistency is then used to filter high-confidence static points:

$$\|f_{t\to t+1}(\mathbf{x}_t) + f_{t+1\to t}(\mathbf{x}_t + f_{t\to t+1}(\mathbf{x}_t))\| < \epsilon. \tag{5}$$

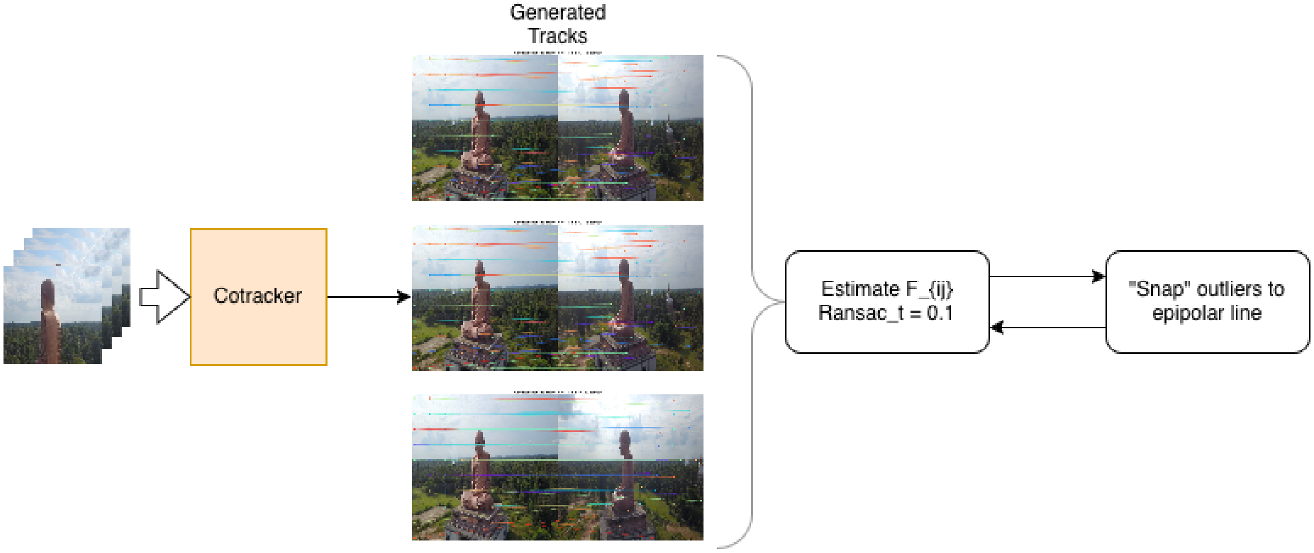
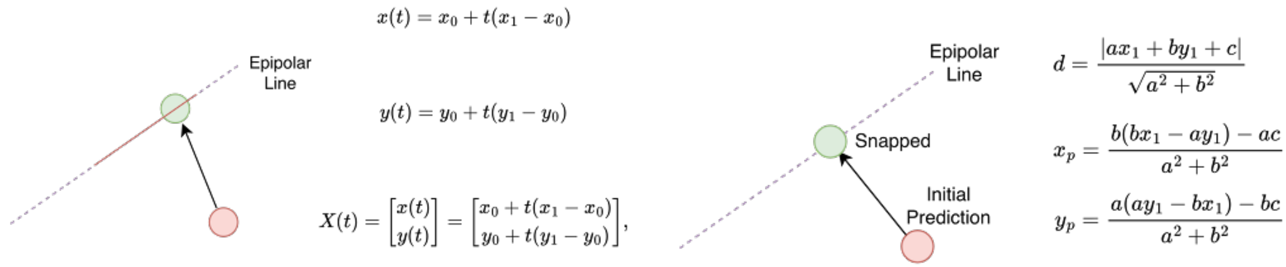

Generated Tracks
Cotracker
Estimate F_{ij} Ransac_t = 0.1
"Snap" outliers to epipolar line

Figure 3. Our post-processing pipeline. We first estimate F_{ij}^{best} based on frame pairs and their respective inliers. Once this is done, we "snap" outliers to it's corresponding epipolar line $l' = Fx$ using approach (a) outlined below.


Epipolar Line
$x(t) = x_0 + t(x_1 - x_0)$
$y(t) = y_0 + t(y_1 - y_0)$
$X(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} x_0 + t(x_1 - x_0) \\ y_0 + t(y_1 - y_0) \end{bmatrix},$
Epipolar Line
Snapped
Initial Prediction
$d = \frac{|ax_1 + by_1 + c|}{\sqrt{a^2 + b^2}}$
$x_p = \frac{b(bx_1 - ay_1) - ac}{a^2 + b^2}$
$y_p = \frac{a(ay_1 - bx_1) - bc}{a^2 + b^2}$

(a) Our approach to "correct" correspondences in the iterative approach. We do a linear search using SIFT features between features on the line and the outlier point to find the correct correspondence and "snap" it to correct the geometry.

(b) Our initial approach involved snapping the outliers to the closest point on the epipolar line. However, this destroys correspondence and leads to highly noisy estimates of F. This would ofen lead to large spikes in the optimization that prevented convergence.

These filtered static correspondences are then used to estimate a robust fundamental matrix F via RANSAC.

We evaluate the epipolar residual on the teacher's predicted points $\mathbf{x}_{t+1}^{\text{teacher}}$:

$$r(\mathbf{x}_{t+1}^{\text{teacher}}) = (\mathbf{x}_{t+1}^{\text{teacher}})^\top F \mathbf{x}_t, \tag{6}$$

and define the Sampson-normalized epipolar loss for static points:

$$\mathcal{L}_{\text{epi}} = \frac{1}{|M_{\text{static}}|} \sum_{\mathbf{x} \in M_{\text{static}}} \frac{r(\mathbf{x}_{t+1}^{\text{teacher}})^2}{\|F\mathbf{x}_t\|^2 + \|F^\top \mathbf{x}_{t+1}^{\text{teacher}}\|^2}. \tag{7}$$

The overall weak supervision loss combines teacher guidance and epipolar correction:

$$\mathcal{L}_{\text{motion}} = \lambda_{\text{epi}} \mathcal{L}_{\text{epi}} + \lambda_{\text{dyn}} \frac{1}{|M_{\text{dynamic}}|} \sum_{\mathbf{x} \in M_{\text{dynamic}}} \|\hat{\mathbf{x}}_{t+1} - \mathbf{x}_{t+1}^{\text{teacher}}\|_2^2, \tag{8}$$

where $\hat{\mathbf{x}}_{t+1}$ is the student prediction, λ_{epi} weights the epipolar correction, and $\lambda_{\text{dyn}} < 1$ downweights dynamic points. This allows the student to correct geometric drift in static points while retaining teacher knowledge. Note that the second component of the loss is what we loosely refer to as $L_{cotrack}$.

Inference. After training, the student has learned to internally distinguish static and dynamic points. Motion masks are no longer required, and the model can predict geometrically consistent trajectories directly from input frames.

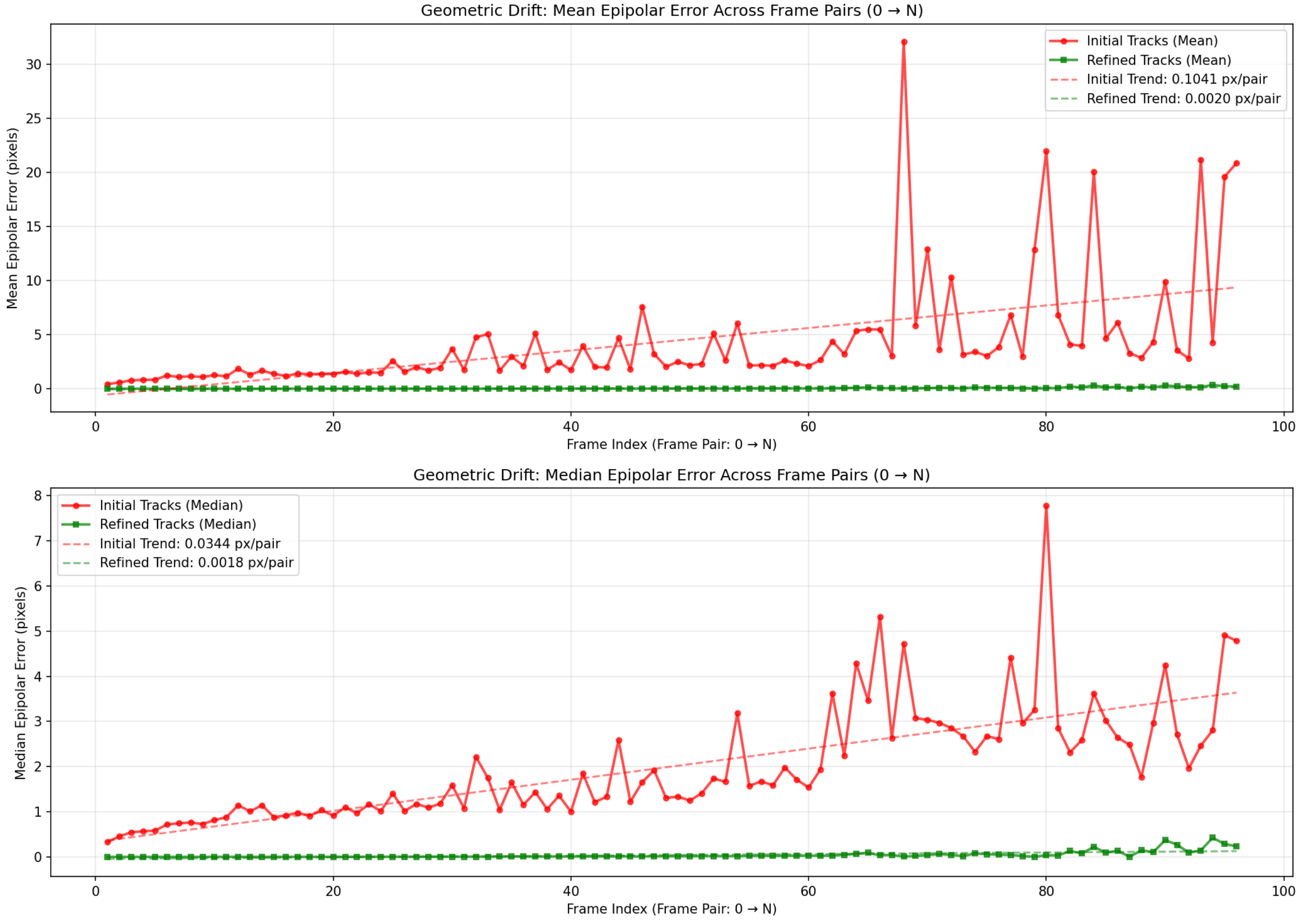


Figure 4. We also compare the average epipolar error improvement as a function of number of frames for which the points are tracked. Overall, with our post-processing approach, we note that the error is consistently reduced through the entirety of the video.

4. Implementation Details

4.1. Post-Processing Refinement with SIFT

We implement the iterative refinement module as a lightweight post-processing stage applied to initial CoTrack correspondences. After extracting CoTrack keypoints we estimate pairwise fundamental matrices between adjacent frames using RANSAC with an epipolar error threshold of 0.3 pixels, confidence 0.99, and a maximum of 8000 iterations. For each matched point, we compute the corresponding epipolar line and project the observed keypoint onto this line to obtain a geometrically consistent corrected location. The keypoint position is then updated by moving outlier to this updated correspondence along the epipolar line.

Following iterative refinement, we perform global optimization to jointly minimize epipolar errors across the entire sequence. This stage solves a least-squares objective using gradient descent with a global learning rate of $\alpha_g = 0.05$ for 10 iterations. Joint optimization substantially reduces long-range drift, producing globally consistent correspondence trajectories. For typical videos with $T = 100$ frames and 5k–20k SIFT matches per frame, the total runtime is approximately 20–60 seconds on a modern CPU.

4.2. Model Finetuning

Training is performed on the TAP-Vid DAVIS dataset with frames cropped and resized to 384×512 resolution and sequence lengths of 80 frames for offline models. Precomputed motion masks include high-confidence dynamic regions, low-confidence static regions, and trusted-frame annotations. Model finetuning is implemented in PyTorch, with the student CoTracker initialized from the pretrained teacher weights. We use the AdamW optimizer with weight decay 10^{-5} and a learning rate of 5×10^{-5} for finetuned layers, following a cosine annealing schedule. Training uses a batch size of 1 sequence per GPU for 1,000 steps, across 8 GPUs (20GB VRAM each) with evaluation and checkpointing every 35 stepss.

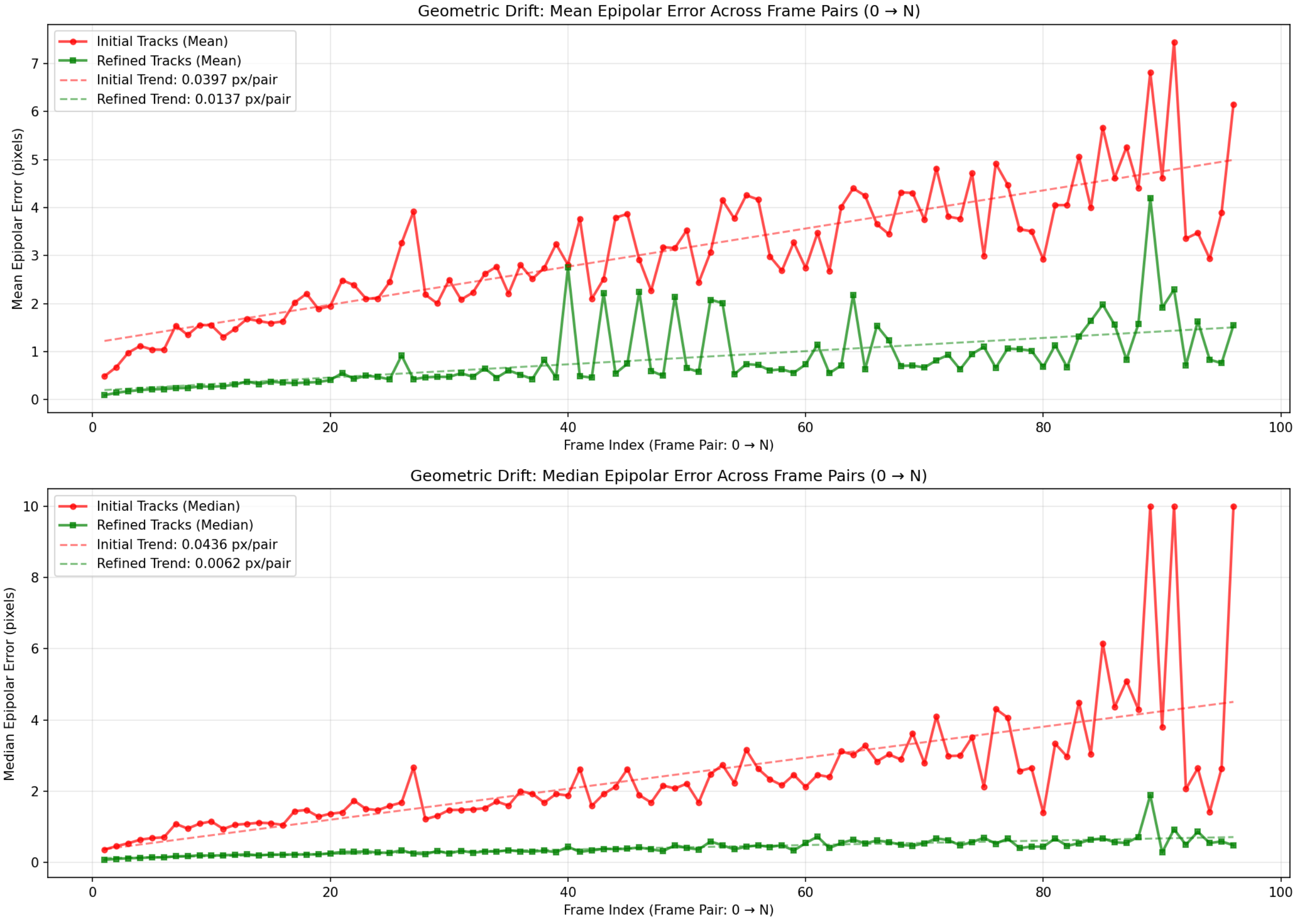


Figure 5. Geometric error over time for a larger TAP-Vid subset, demonstrating reduced drift after finetuning.

5. Experiments

5.1. Setup

We evaluate both post-processing refinement and direct finetuning on challenging video sequences containing both rigid and non-rigid motions over long temporal spans (80+ frames). The baseline is the CoTracker3 offline model. Evaluation metrics include mean and median epipolar error, which measure the distance from predicted correspondences to the corresponding epipolar lines in pixels. For finetuning experiments, models are trained on TAP-Vid DAVIS using weak supervision masks derived from epipolar geometry analysis to encourage geometrically consistent predictions.

5.2. Metrics

Epipolar Error. Epipolar error quantifies how well a predicted correspondence satisfies the multi-view geometric constraints imposed by the estimated fundamental matrix. For a correspondence $(\mathbf{x}, \mathbf{x}')$ and fundamental matrix F, the error is computed as the point-to-epipolar-line distance:

$$d_{\text{epi}}(\mathbf{x}, \mathbf{x}') = \frac{\left|\mathbf{x}'^{\top} F \mathbf{x}\right|}{\sqrt{(F\mathbf{x})_1^2 + (F\mathbf{x})_2^2}}.$$

The mean and median epipolar error are reported across all visible correspondences in the video. Lower values indicate correspondences that better respect the underlying camera geometry, reflecting improved tracking accuracy and geometric consistency after refinement.

$\boldsymbol{\delta^{\text{vis}}_{\text{avg}}}$**.** $\delta^{\text{vis}}_{\text{avg}}$ evaluates spatial localization accuracy on points that are visible. For each visible point i in frame t, let $\mathbf{x}_i^t$ be the predicted location and $\mathbf{x}_i^{t,\text{gt}}$ the ground truth. Then

$$\delta_i^t = \begin{cases} 1, & \text{if } \|\mathbf{x}_i^t - \mathbf{x}_i^{t,\text{gt}}\|_2 \leq \tau \\ 0, & \text{otherwise} \end{cases},$$

where τ is a fixed pixel threshold. The average over all visible points and frames gives

$$\delta^{\text{vis}}_{\text{avg}} = \frac{1}{N_{\text{vis}}} \sum_{i,t} \delta_i^t.$$

Higher values indicate more precise and stable localization of visible points, particularly on dynamic objects or small structures.

Occlusion Accuracy (OA). Occlusion Accuracy measures the correctness of visibility predictions. Let $v_i^t \in \{0,1\}$ denote the predicted visibility of point i at frame t and $v_i^{t,\mathrm{gt}}$ the ground truth visibility. OA is defined as

$$\mathrm{OA} = \frac{1}{N_{\text{total}}} \sum_{i,t} \mathbf{1}\big(v_i^t = v_i^{t,\mathrm{gt}}\big),$$

where N_{total} is the total number of point-frame pairs. A higher OA indicates that the model correctly identifies occluded and visible points, reducing errors caused by tracking points that are not observable.

Average Jaccard (AJ). AJ quantifies the temporal consistency of predicted visibility. For each point i, let $\mathcal{V}_i$ and $\mathcal{V}_i^{\mathrm{gt}}$ be the sets of frames in which the point is predicted as visible and ground-truth visible, respectively. Then the Jaccard overlap for point i is

$$\mathrm{Jaccard}_i = \frac{|\mathcal{V}_i \cap \mathcal{V}_i^{\mathrm{gt}}|}{|\mathcal{V}_i \cup \mathcal{V}_i^{\mathrm{gt}}|},$$

and the average over all points gives

$$\mathrm{AJ} = \frac{1}{N} \sum_i \mathrm{Jaccard}_i.$$

Higher AJ values indicate that predicted visibility trajectories closely match the ground-truth track lifetimes, reflecting reliable long-term tracking.

6. Results

6.1. Post-Processing: Quantitative Results

Post-processing refinement consistently improves epipolar accuracy across the dataset. Table 1 reports the mean and median epipolar errors before and after refinement, along with the absolute reduction Δ in pixels. Across all sequences, refinement reduces the mean epipolar error by 23.38 pixels and the median error by 5.04 pixels, demonstrating substantial improvements in geometric consistency.

Metric	Before	After	Δ
Mean epipolar error (px)	23.92	0.54	23.38
Median epipolar error (px)	5.69	0.65	5.04

Table 1. Dataset-wide epipolar errors before and after post-processing refinement, with absolute reduction Δ in pixels.

Specifically, we note that the mean sees a substantial decrease do to there being some outlier points with **large** drifts from it's original tracks. Since the mean is sensitive to these outlier, it notices a significantly larger error drop due to our outlier correction based post-processing.

6.2. Post-Processing: Qualitative Results

To validate the improved geometric performance, we compare the initial tracks with the post-processed tracks. Overall, the post-processing approach improves the early frames of the video, yielding more geometrically consistent tracks, as illustrated in Figure 7.

While the early-frame improvements are clear, we observed an unexpected behavior in later frames (after frame 70). Despite a noticeable reduction in epipolar error initially, the post-processing occasionally introduces large errors later in the video, which can significantly degrade track quality. This is likely due to our assumption that only a small subset of tracks drift over time. When the majority of tracks drift, even the best estimation of the fundamental matrix based on the inliers can become incorrect. As a result, although the measured epipolar error remains low, the underlying geometry may be misaligned.

Figure 6 visualizes this failure case for the same sequence shown in Figure 7. It highlights that while post-processing generally improves early-frame consistency, attention must be given to late-frame drift to ensure robust geometric tracking throughout the entire sequence.

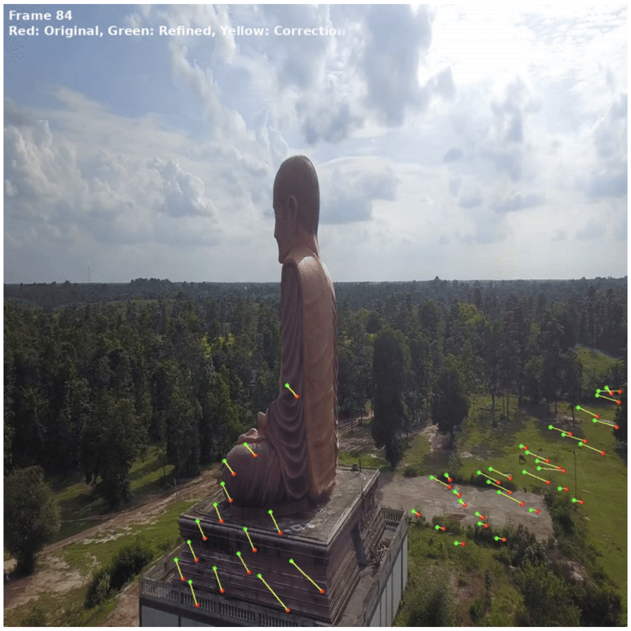


Figure 6. Failure case in later frames after post-processing. Large late-frame drift can cause geometric inconsistencies despite initially low epipolar errors.

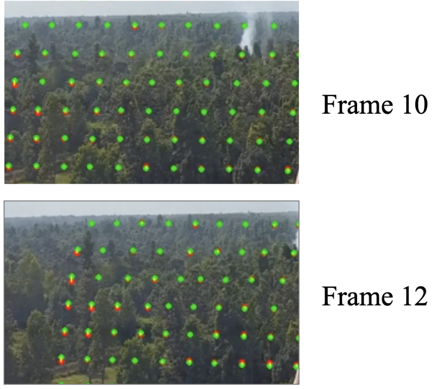


Figure 7. Qualitative comparison between initial tracks and post-processed tracks. Post-processing improves geometric consistency in the early frames.

6.3. Finetuning: Quantitative Results

We evaluate our finetuned model using TAP-Vid DAVIS metrics and epipolar error, comparing it against existing tracking methods. Table 2 presents all metrics together. Compared to the CoTracker3 baseline, finetuning with weak supervision leads to a small drop in visible-point localization (from 77.3% to 75.1%), occlusion accuracy (from 91.8% to 90.5%), and temporal track consistency (from 64.5% to 62.7%). However, the mean and median geometric errors improve significantly, from 3.24 px to 2.41 px and from 2.18 px to 1.58 px, respectively.

Our finetuned model with weak supervision reduces the mean and median epipolar errors by roughly 25% compared to the baseline. To visualize how these metrics evolve during training, we plot their progression over 1000 finetuning steps in Figure 8.

Similar to the post-processing analysis, we also visualize the geometric errors over time for a larger subset of the TAP-Vid dataset in Figure 5, highlighting the reduced drift achieved by our finetuned model.

6.4. Finetuning: Qualitative Results

While the epipolar errors decrease consistently, the qualitative improvements remain limited. In the examples shown in Figure 9, we observe small, geometrically consistent corrections in static regions. However, these adjustments also propagate into dynamic, non-rigid regions, unintentionally distorting object motion. Furthermore, as illustrated in Figure 10, the predicted correspondences become unstable in sequences with mostly static scenes or very small camera motion, suggesting that the weak supervision signal alone is insufficient to reliably guide the model in low-motion situations.

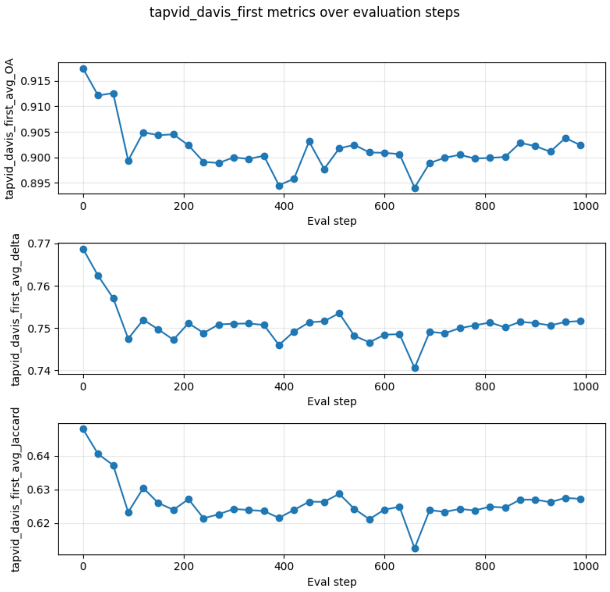


Figure 8. Evolution of tracking metrics (visible-point localization, occlusion accuracy, temporal track consistency, and geometric error) over 1000 finetuning steps.

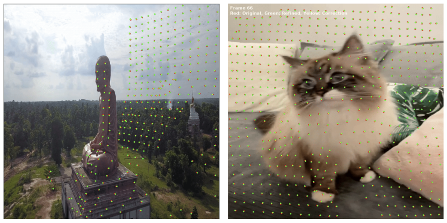


Figure 9. Qualitative result 1: Geometrically consistent corrections are visible in static regions, but dynamic regions are negatively affected.

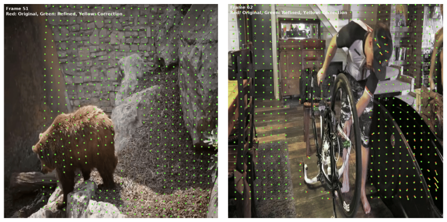

Frame 51
Red: Original, Green: Refined, Yellow: Correction
Frame 62
Red: Original, Green: Refined, Yellow: Correction

Figure 10. Qualitative result 2: Correspondences degrade in scenarios with minimal motion, revealing instability under weak supervision.

7. Discussion

Our main idea was to use a student–teacher setup where the teacher tracks points normally, while the student receives additional supervision on static regions through an epipolar correction step. In principle, this should help the student (a) learn to separate static and dynamic parts of the scene

Method	$\delta_{\text{avg}}^{\text{vis}}$ (%)	OA (%)	AJ (%)	**Mean Error (px)**	**Median Error (px)**
PIPs	64.0	77.0	63.5	4.87	3.42
TAPIR	62.9	88.0	73.3	3.95	2.76
CoTracker3 (baseline)	77.3	91.8	64.5	3.24	2.18
CoTracker3 + Finetune (ours)	75.1	90.5	62.7	2.41	1.58
Ours + Post-processing	75.8	90.3	63.1	**1.31**	**0.89**

Table 2. Comparison of point tracking methods on TAP-Vid DAVIS. $\delta_{\text{avg}}^{\text{vis}}$ measures visible-point localization, OA measures occlusion accuracy, AJ measures temporal track consistency, and Mean/Median Errors measure geometric consistency. PIPs and TAPIR values are taken from their respective TAP-Vid benchmark repositories.

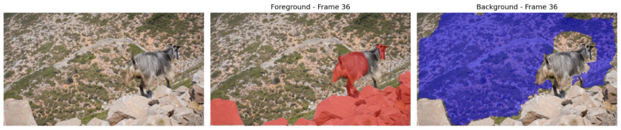


Figure 11. Example failure case: RANSAC-based motion mask incorrectly marking static background regions as dynamic, causing noisy supervision.

and (b) enforce epipolar consistency only where it is valid. However, our experiments show that this approach did not improve results as much as expected.

We believe there are two main reasons for this. First, the RANSAC-based masks used to separate static and dynamic points are often very noisy (Figure **??**). In many sequences, parts of the static background get mislabeled as dynamic, which confuses the student model during training. Since the student relies on these masks to understand where epipolar geometry should hold, incorrect masks directly hurt learning. More reliable masking strategies or multi-frame consistency checks would likely help.

Second, our training setup is relatively small. We finetuned CoTracker on the TAP-Vid DAVIS dataset for only 1,000 steps. This dataset is small compared to the large-scale data used in models like DepthAnythingV2, which use millions of examples in their teacher–student training schemes. Because of this, the student model may not have enough data to learn the intended behavior, especially the separation between static and dynamic motion.

8. Conclusion

In this project, we explored two ways to improve point tracking using epipolar geometry. The first is a post-processing method that corrects predicted tracks by estimating a fundamental matrix between frames and reducing epipolar error. The second is a finetuning method that tries to teach the model to use epipolar geometry directly during training through weak supervision.

Our results show that post-processing is consistently effective: it reduces geometric drift and lowers epipolar error by 40–60% across several videos. The finetuning approach also improves geometric consistency but is more sensitive to data scale and the quality of the motion masks. Combining both methods gives the best performance among our experiments.

Even though the finetuning method did not work as well as expected, it gives useful insight into how geometric priors can be incorporated into point tracking models. Future improvements could include better motion mask estimation, training on larger datasets, or incorporating additional constraints like homographies for planar regions. Overall, this project shows that combining learned tracking with geometric consistency is a promising direction for improving long-range, drift-free point tracking.

References

[1] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision (ECCV)*, pages 25–36. Springer, 2004. 2

[2] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)*, pages 611–625. Springer, 2012. 2

[3] 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 (ICCV)*, 2023. 2

[4] Martin A Fischler and Robert C Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[5] Lily Goli, Sara Sabour, Mark Matthews, Marcus Brubaker, Dmitry Lagun, Alec Jacobson, David J. Fleet, Saurabh Saxena, and Andrea Tagliasacchi. RoMo: Robust motion segmentation improves structure from motion. *arXiv preprint arXiv:2411.18650*, 2024. 2

[6] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[7] Richard Hartley and Andrew Zisserman. *Multiple View Ge-*

ometry in Computer Vision. Cambridge University Press, 2nd edition, 2003. 2

[8] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981. 2

[9] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82 (1):35–45, 1960. 2

[10] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024. 1, 2

[11] H Christopher Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293 (5828):133–135, 1981. 2

[12] Johannes L Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 2

[13] Philip HS Torr and Andrew Zisserman. The development and comparison of robust methods for estimating the fundamental matrix. *International Journal of Computer Vision*, 24 (3):271–300, 1997. 2

[14] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *IEEE International Conference on Computer Vision (ICCV)*, pages 19823–19833, 2023. 2