You Are Your Best Teacher: Semi-Supervised Surgical Point Tracking with Cycle-Consistent Self-Distillation

Anonymous CVPR submission

Paper ID 60

Abstract

Synthetic datasets have enabled significant progress in point tracking by providing large-scale, densely annotated supervision. However, deploying these models in real-world domains remains challenging due to domain shift and lack of labeled data—issues that are especially severe in surgical videos, where scenes exhibit complex tissue deformation, occlusion, and lighting variation. While recent approaches adapt synthetic-trained trackers to natural videos using teacher ensembles or augmentation-heavy pseudo-labeling pipelines, their effectiveness in high-shift domains like surgery remains unexplored. This work presents SurgTracker, a semi-supervised framework for adapting synthetic-trained point trackers to surgical video using filtered self-distillation. Pseudo-labels are generated online by a fixed teacher—identical in architecture and initialization to the student—and are filtered using a cycle consistency constraint to discard temporally inconsistent trajectories. This simple yet effective design enforces geometric consistency and provides stable supervision throughout training, without the computational overhead of maintaining multiple teachers. Experiments on the STIR benchmark show that SurgTracker improves tracking performance using only 80 unlabeled videos, demonstrating its potential for robust adaptation in high-shift, data-scarce domains.

1. Introduction

Tracking visual points over time is a core problem in computer vision, underpinning applications in motion understanding, visual correspondence, and robotic perception. Recent advances in learning-based point trackers [2, 4, 9, 10] have shown remarkable performance by training on large-scale synthetic datasets with dense supervision. These models benefit from scalability and control in simulation, but transferring them to real-world scenarios remains a major challenge due to domain shift and lack of annotated data.

To mitigate this gap, recent efforts [5, 8] propose semi-supervised adaptation strategies using pseudo-labels generated on unlabeled natural videos. These methods leverage teacher-student frameworks and consistency losses to refine models in the absence of ground truth. However, they have been validated primarily on natural video domains, which, despite being unlabeled, still resemble the synthetic training distribution in terms of motion regularity and scene composition. Their applicability to more specialized, high-variance domains remains largely unexplored.

One such domain is surgical video analysis, where accurate point tracking can facilitate understanding of tissue dynamics, tool-tissue interaction, and intraoperative state estimation—critical for applications such as surgical skill assessment, automation, and guidance [13]. However, the domain poses unique challenges: deformable anatomy, specular lighting, heavy occlusion, and rapid motion. Moreover, obtaining annotated datasets for point tracking in surgery is impractical due to privacy concerns, the need for domain expertise, and the high cost of manual labeling.

Prior methods in point tracking in surgical videos have typically relied on classical techniques such as sparse feature matching or optical flow [7]. Recent work such as SurgMotion [18] adapts OmniMotion [17] to surgical data using domain-specific priors, but requires test-time optimization, making it less practical for real-time deployment. As a result, the question remains: *can recent synthetic-trained point trackers be effectively adapted to surgical video—without any manual annotations?*

To address this, we propose SurgTracker, a semi-supervised framework for adapting synthetic-trained point trackers to surgical video using only unlabeled data. While CoTracker3 [8] adapts to natural videos using pseudo-labels from diverse teacher models, we find that this approach is less effective in surgical settings, where the domain shift is more pronounced. Instead, SurgTracker employs a simpler yet more effective strategy: it leverages pseudo-labels from a single frozen teacher, identical to the student in architecture and initialization, and applies a cycle consistency constraint to retain only temporally coherent trajectories.

We attribute effectiveness of this design to three factors:

first, diverse teachers introduce higher supervision variance due to inconsistent behaviors under domain shift, making pseudo-label quality less reliable; second, architectural alignment between teacher and student improves representational compatibility, allowing for more effective learning; and third, using a fixed teacher yields a stable supervisory signal across training batches, reducing fluctuations in optimization dynamics. In addition, our single-teacher setup eliminates the need to keep multiple large models in memory during training, making the approach more computationally efficient. Experiments on STIR benchmark [18] show that SurgTracker improves tracking performance using only 80 unlabeled videos, demonstrating that in high-shift data-scarce domains, supervision consistency and alignment can outweigh benefits of teacher diversity.

2. Related Works

2.1. Point Tracking

Deep learning-based point trackers have advanced rapidly, largely by training on synthetic datasets due to the difficulty of labeling real-world trajectories. Early work like PIPs [6] framed dense tracking as long-range motion estimation, later extended to longer sequences in PIPs++ [19]. TAPIR [4] built on this by introducing global matching, while CoTracker [9] leveraged transformers to jointly track multiple points and better handle occlusion. More recent variants like LocoTrack [2] uses 4D correlation volumes whereas Track-On [1] enables frame-by-frame tracking using spatial and context memory. While these methods show strong performance, they are trained on synthetic datasets and have been validated primarily on natural video domains.

Point tracking in surgical videos is essential for modeling tissue dynamics and enabling image-guided robotic interventions [18]. Classical methods based on sparse features or dense optical flow [7] are limited by poor texture, deformation, and occlusion in surgical scenes. Recent approaches such as SENDD [12] use graph-based models to jointly estimate 2D correspondences and 3D deformation. More recently, Zhan et al. [18] introduced a benchmark with manually annotated trajectories and proposed SurgMotion, which adapts OmniMotion [17] with domain-specific priors. While effective, SurgMotion relies on test-time optimization, limiting its applicability in real-time settings. In contrast, our work explores whether synthetic-trained trackers can be adapted to surgical videos without any labels to enable robust real-time performance in clinical scenarios.

2.2. Unsupervised Domain Adaptation

While synthetic data enables scalable training, domain shift remains a core challenge when deploying models on real-world videos. Self-training with pseudo-labels has emerged as a promising strategy, wherein source-trained models generate labels on unlabeled target data to guide fine-tuning. BootsTAP [5] applies this paradigm to large-scale natural video via teacher-student learning and strong augmentations. CoTracker3 [8] improves efficiency by distilling pseudo-labels from multiple teacher models, but applies no filtering to account for label noise. Sun et al. [14] incorporate cycle consistency to improve label quality, but compute pseudo-labels only once and keep them fixed, increasing susceptibility to confirmation bias.

Critically, these approaches have been validated only on natural videos, and it remains unclear whether they generalize to domains with significantly higher distribution shift—like surgical video. We address this gap by extending self-training-based point tracking to surgical data, leveraging a single, architecture-aligned teacher and applying cycle consistency filtering to provide stable supervision.

3. Method

3.1. Problem Formulation

Tracking tissue motion in surgical videos involves accurately following specific tissue points across frames. Given a video sequence $V = \{I_t\}_{t=1}^{T}$ consisting of T frames, our objective is to track a set of N query points $Q = \{(x_i, y_i, t_0)\}_{i=1}^{N}$ where (x_i, y_i) denotes spatial location of i-th query point in the frame t_0. The goal is to estimate a trajectory $P = \{(x_i^t, y_i^t)\}_{t=1}^{T}$ for each query point i, representing its predicted location in every frame of the sequence.

3.2. SurgTracker

We propose SurgTracker, a semi-supervised framework for adapting synthetic-pretrained point trackers to surgical video, where large domain shift and lack of annotations present significant challenges. Our method leverages CoTracker3—pretrained on synthetic data and adapted to natural videos—as a fixed teacher to produce pseudo-labels, which are then filtered via a cycle consistency constraint to remove noisy trajectories. The student model, identical in architecture and initialization to the teacher, is then fine-tuned using these filtered labels. An overview of the SurgTracker pipeline is shown in Fig. 1.

Unlike prior work that relies on teacher model ensembles [8] or large-scale data augmentation [5], SurgTracker uses a single teacher—architecturally aligned with the student—and leverages temporal consistency to identify high-quality training signals. This simple yet effective design enables adaptation to surgical videos without requiring any annotations. The method consists of three main stages: (1) pseudo-label generation, (2) trajectory filtering via cycle consistency, and (3) supervised fine-tuning of the student.

3.2.1. Pseudo-Label Generation

For each training sequence, we sample a set of query points Q from the first frame. To ensure that these points are infor-

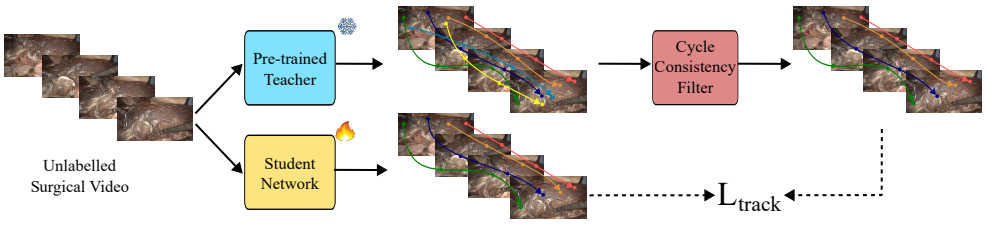

Figure 1. Overview of the SurgTracker framework. Given an unlabeled surgical video, pseudo-labels are generated by a frozen teacher network and filtered using a cycle consistency check to remove temporally inconsistent trajectories. The filtered trajectories supervise the student model, which is fine-tuned using a tracking loss $\mathcal{L}_{\text{track}}$. The teacher model remains frozen during training.

mative and trackable, we extract keypoints using SIFT [11], which provides robust features under appearance changes and viewpoint variation. Sequences with an insufficient number of detected keypoints are excluded to maintain supervision quality. The teacher model M then predicts candidate trajectories $\hat{P}$ for each query point $q_i \in Q$.

3.2.2. Cycle-Consistent Filtering

To improve the quality of pseudo-labels, we apply a cycle consistency check to identify and discard noisy trajectories. Let $\hat{P}_i = \{(x_i^t, y_i^t)\}_{t=t_0}^{t_1}$ denote the forward trajectory for a query point $q_i = (x_i^{t_0}, y_i^{t_0})$ generated by the teacher M, where t_0 and t_1 are the start and end frames of the sequence. We then perform reverse tracking starting from the final predicted location $(x_i^{t_1}, y_i^{t_1})$, obtaining a backward trajectory $\tilde{P}_i = \{(\tilde{x}_i^t, \tilde{y}_i^t)\}_{t=t_1}^{t_0}$, again using M. We define cycle consistency error as the Euclidean distance between original query point and endpoint of the backward track:

$$\mathcal{E}_{\text{cycle}}(q_i) = \left\| (x_i^{t_0}, y_i^{t_0}) - (\tilde{x}_i^{t_0}, \tilde{y}_i^{t_0}) \right\|_2 . \quad (1)$$

A trajectory $\hat{P}_i$ is considered valid if the cycle consistency error satisfies $\mathcal{E}_{\text{cycle}}(q_i) < \alpha$, where α is a hyperparameter controlling the filtering aggressiveness. Only valid trajectories are used as pseudo-labels to supervise the student. Throughout training, the teacher model is frozen and only the student model is updated via backpropagation.

3.2.3. Student Fine-Tuning

We train the student using supervision from both visible and occluded trajectories, following the loss formulation in CoTracker3 [8]. Tracking supervision is provided via a Huber loss with a threshold of 6, applied across multiple refinement iterations. To emphasize visible points, higher weight is assigned to their loss terms, while occluded points are down-weighted by a factor of 1/5. An exponential discount factor $\gamma \in (0, 1)$ is also applied, reducing contribution of earlier iterations and encouraging accurate predictions in final refinement steps. The overall loss is defined as:

$$\mathcal{L}_{\text{track}} = \sum_{k=1}^{K} \gamma^{K-k} \left(\mathbf{1}_{\text{vis}} + \frac{1}{5} \mathbf{1}_{\text{occ}} \right) \cdot \text{Huber}(\mathcal{P}^{(k)}, \mathcal{P}^{\star}) \quad (2)$$

where $\mathcal{P}^{(k)}$ is the student's prediction at refinement iteration k, and $\mathcal{P}^{\star}$ is the pseudo-label provided by teacher M. Since pseudo-labels can be noisy, we found it more stable to omit confidence and visibility supervision during fine-tuning. This helps prevent overfitting to unreliable label quality and focuses learning on trajectory refinement.

4. Experiments

4.1. Datasets and Metrics

We train on the Cholec80 dataset [16], which contains 80 laparoscopic cholecystectomy videos exhibiting diverse anatomy, motion patterns, lighting conditions, and tool interactions. The videos are recorded at 25 FPS with an average duration of 2,306 seconds. Although it lacks point-level annotations, we use it as an unlabeled dataset for semi-supervised training. For evaluation, we use the STIR benchmark [18], comprising around 425 in-vivo and ex-vivo surgical videos recorded with a da Vinci Xi robot and annotated with over 3,000 points in the first and last frames of each sequence. We filter out around 20 sequences with excessive label noise to ensure consistent evaluation.

We evaluate tracking performance using three metrics: Mean Endpoint Error (MEE), Mean Chamfer Distance (MCD), and Average Accuracy $< \delta^x_{avg}$, as defined in TAP-Vid [3]. The $< \delta^x_{avg}$ metric is computed as the average percentage of tracked points falling within thresholds of {4, 8, 16, 32, 64} pixels from ground truth positions.

4.2. Implementation Details

The student model is trained for 120,000 iterations using the Adam optimizer with a cosine learning rate schedule starting at 5×10^{-5}. Each batch contains a randomly sampled sequence with 64 query points tracked over 16 frames, sampled with a random stride between 1 and 4. Training is conducted on NVIDIA RTX 4090 GPUs. The cycle consistency threshold $\alpha = 5$ provides the best trade-off between label quality and training signal.

4.3. Results

We evaluate SurgTracker on the STIR benchmark, comparing it to several recent methods for point tracking, in-

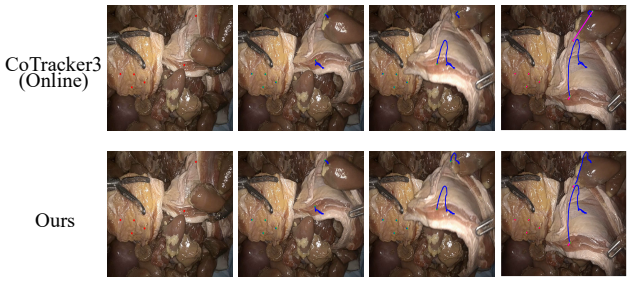

Figure 2. Comparison of CoTracker3 and our method on a challenging sequence. Red and green dots mark initial and mid-frame predicted positions, respectively, blue lines show trajectories, and pink lines indicate final error. Our model better handles occlusion and motion change, accurately recovering the original trajectory.

Table 1. Performance comparison on the STIR dataset.

Method	MEE ↓	MCD ↓	$< \delta^x_{avg}$ ↑
RAFT	44.25	43.60	50.41
SENDD	22.80	45.18	66.5
TAPIR	24.33	25.03	61.04
BootsTAP	20.38	21.4	63.74
CoTracker3 (Online)	17.01	17.81	68.11
SurgTracker (Ours)	**16.27**	**17.12**	**68.55**

cluding RAFT [15], SENDD [12], TAPIR [4], BootsTAP [5], and CoTracker3 (Online) [8]. As shown in Table 1, SurgTracker outperforms all baselines across all metrics. Compared to CoTracker3 (Online), which serves as our initialization and frozen teacher, it reduces MEE by 0.74 and MCD by 0.69 while improving $< \delta^x_{avg}$ by 0.44. These gains demonstrate the effectiveness of filtered self-distillation for adapting point trackers to high-shift surgical domains.

Figure 2 shows a qualitative comparison with CoTracker3 on a challenging occlusion scenario. Both models initially track the top-right point correctly until an occlusion occurs (second column). Notably, during the occlusion, the direction of the intended motion changes. While CoTracker3 drifts and continues tracking the occluding tissue with an estimation of the prior motion, our model successfully recovers and resumes tracking the original structure after it reappears (third column). The final trajectory is significantly more accurate, highlighting the robustness of our distilled model to occlusions and motion changes.

4.4. Ablation Studies

To evaluate the impact of cycle consistency filtering, we vary the threshold α controlling the maximum allowed deviation between a point and its cycle-tracked counterpart. As shown in Table 2, omitting the filter results in lower accuracy, confirming the presence of noisy pseudo-labels. Filtering with $\alpha = 5$ achieves the best trade-off, minimizing both MEE and MCD while improving $< \delta^x_{avg}$, reflecting more accurate tracking. A lower threshold (α = 2.5) is overly conservative, discarding too many training samples and thus limiting supervision. Conversely, a higher threshold (α = 7.5) allows more trajectories but admits additional noise, slightly degrading performance. These results highlight the importance of temporal consistency in improving label quality and overall tracking performance.

Table 2. Ablation on cycle consistency threshold α.

α	MEE ↓	MCD ↓	$< \delta^x_{avg}$ ↑
No filtering	16.69	17.46	68.04
2.5	16.76	17.58	68.02
5	**16.27**	**17.12**	**68.55**
7.5	16.43	17.23	68.31

Table 3. Comparison of different teacher configurations. The student is always CoTracker3 (Online). CoT3 (On) and CoT3 (Off) refer to online and offline versions of CoTracker3 respectively.

Teacher Models	MEE ↓	MCD ↓	$< \delta^x_{avg}$ ↑
CoT3 (On)	**16.27**	**17.12**	**68.55**
CoT3 (On), CoT3 (Off)	16.28	17.10	68.50
CoT3 (On), CoT3 (Off), BootsTAP	16.38	17.21	68.39
CoT3 (On), CoT3 (Off), Track-On	16.80	17.64	68.00

Table 3 compares different teacher configurations for pseudo-label generation, following multi-teacher setup in CoTracker3. For each configuration, a teacher model is randomly sampled from corresponding pool per batch and used to generate pseudo-labels for student fine-tuning. Our self-distillation approach, which uses only CoTracker3 (Online) as the teacher, achieves the best performance. Incorporating other teachers slightly degrades performance, likely due to inconsistent supervision that hinders stable learning. These findings suggest that, under a significant domain shift, a consistent, architecture-aligned teacher can outperform diverse ensembles, offering more effective supervision.

4.5. Conclusion

We present SurgTracker, a semi-supervised framework for adapting synthetic-trained point trackers to surgical video through filtered self-distillation. By leveraging a single, architecture-aligned teacher and enforcing cycle consistency, our method provides stable, high-quality supervision without the overhead of maintaining teacher ensembles. Experiments on the STIR benchmark demonstrate that SurgTracker improves tracking performance using only 80 unlabeled videos, demonstrating that consistent supervision can outperform diverse teacher setups in challenging, high-shift surgical domains.

References

[1] Görkay Aydemir, Xiongyi Cai, Weidi Xie, and Fatma Güney. Track-on: Transformer-based online point tracking with memory. *arXiv preprint arXiv:2501.18487*, 2025. 2

[2] 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, 2

[3] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022. 3

[4] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10061–10072, 2023. 1, 2, 4

[5] Carl Doersch, Pauline Luc, Yi Yang, Dilara Gokay, Skanda Koppula, Ankush Gupta, Joseph Heyward, Ignacio Rocco, Ross Goroshin, João Carreira, et al. Bootstap: Bootstrapped training for tracking-any-point. In *Proceedings of the Asian Conference on Computer Vision*, pages 3257–3274, 2024. 1, 2, 4

[6] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision*, pages 59–75. Springer, 2022. 2

[7] Sontje Ihler, Max-Heinrich Laves, and Tobias Ortmaier. Patient-specific domain adaptation for fast optical flow based on teacher-student knowledge transfer. *arXiv preprint arXiv:2007.04928*, 2020. 1, 2

[8] 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, 3, 4

[9] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European Conference on Computer Vision*, pages 18–35. Springer, 2024. 1, 2

[10] Hongyang Li, Hao Zhang, Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, and Lei Zhang. Taptr: Tracking any point with transformers as detection. In *European Conference on Computer Vision*, pages 57–75. Springer, 2024. 1

[11] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, pages 1150–1157. Ieee, 1999. 3

[12] Adam Schmidt, Omid Mohareri, Simon DiMaio, and Septimiu E Salcudean. Sendd: Sparse efficient neural depth and deformation for tissue tracking. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 238–248. Springer, 2023. 2, 4

[13] Adam Schmidt, Omid Mohareri, Simon DiMaio, Michael C Yip, and Septimiu E Salcudean. Tracking and mapping in medical computer vision: A review. *Medical Image Analysis*, page 103131, 2024. 1

[14] Xinglong Sun, Adam W Harley, and Leonidas J Guibas. Refining pre-trained motion models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4932–4938. IEEE, 2024. 2

[15] 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. 4

[16] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36 (1):86–97, 2016. 3

[17] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19795–19806, 2023. 1, 2

[18] Bohan Zhan, Wang Zhao, Yi Fang, Bo Du, Francisco Vasconcelos, Danail Stoyanov, Daniel S Elson, and Baoru Huang. Tracking everything in robotic-assisted surgery. *arXiv preprint arXiv:2409.19821*, 2024. 1, 2, 3

[19] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19855–19865, 2023. 2