"Box it and Track it: A Weakly Supervised Framework for Cell Tracking"

Nabeel Khalid[1,2*][0000−0001−9274−3757]⋆, Mohammadmahdi Koochali[1*][0000−0001−8780−253X], Khola Naseem[1,2][0000−0003−4785−2588], Gillian Lovell[4][0009−0004−5180−9704], Bianca Migliori[6][0000−0002−6333−314X], Daniel A Porto[6][0000−0002−1021−2467], Johan Trygg[3,7][0000−0002−4239−6520], Andreas Dengel[1,2][0000−0002−6100−8255], and Sheraz Ahmed[1][0000−0002−4239−6520]

[1] German Research Center for Artificial Intelligence (DFKI) GmbH, Kaiserslautern 67663, Germany
firstname.lastname@dfki.de
[2] RPTU Kaiserslautern–Landau, Kaiserslautern 67663, Germany
[3] Sartorius Corporate Research, Sweden
[4] Sartorius, BioAnalytics, Royston, United Kingdom
[5] Sartorius, Digital Solutions, Royston, United Kingdom
[6] Sartorius, BioAnalytics, Ann Arbor, United States
firstname.lastname@sartorius.com
[7] Computational Life Science Cluster (CLiC), Umeå University, Sweden

Abstract. Accurate cell tracking in microscopy is essential for studying biological dynamics like proliferation and migration. Traditional fully supervised methods demand dense pixel-wise masks for every frame, making them impractical for large-scale use. Recent methods like SAT reduce annotation effort by using sparse point-based supervision, but still require multiple positive and negative points per cell, which remains labor-intensive. BoxTrack offers a lightweight and annotation-efficient alternative, requiring only a single bounding box per cell in the first frame. Without relying on any point-level annotations, it performs end-to-end instance segmentation and tracking over entire sequences. This simplification leads to a substantial reduction in annotation cost while improving performance over SAT. On the CTMC dataset, BoxTrack improves Multiple Object Tracking Accuracy (MOTA) by +**15.96%** over SAT. For the CTC dataset, it yields a +**8.86%** MOTA gain. Code is available at https://github.com/nabeelkhalid92/Box-it-Track-it.

Keywords: Microscopy, Cell Tracking, Segment Anything, Weak Supervision, Temporal Downsampling, Deep Learning

1 Introduction

Accurate cell tracking in microscopy is essential for understanding dynamic biological processes such as proliferation, migration, and cell-cell interactions [19,

⋆ These authors contributed equally to this work.

20, 28]. It enables researchers to quantify cellular behaviors over time and plays a central role in studying wound healing, cancer metastasis, immune response, and drug screening [5, 19]. Traditional cell tracking pipelines are heavily based on fully supervised instance segmentation frameworks [5, 12, 24, 26, 28], requiring dense pixel-wise annotations for every frame in a sequence. However, generating such exhaustive labels is prohibitively expensive, especially for high-throughput imaging settings where manual annotation becomes the bottleneck. This has motivated the emergence of weakly supervised approaches [4, 10, 27] that aim to reduce annotation costs without sacrificing tracking performance.

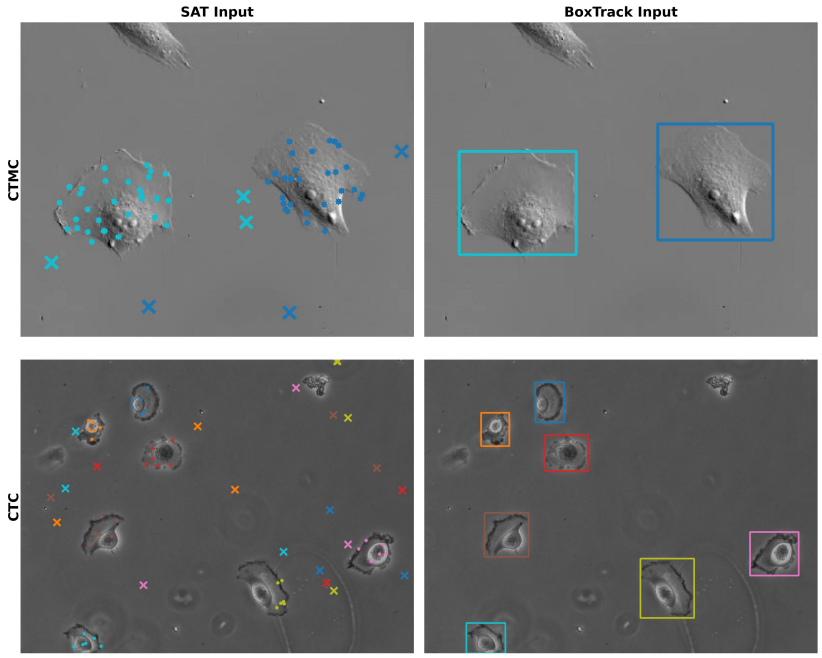

Fig. 1. Comparison of annotation inputs required by SAT and BoxTrack on the first frame of sequences from the CTMC and CTC datasets. SAT relies on dense point-level supervision: 30 positive points (•) and 3 negative points (×) per cell for CTMC, and 6 positive (•) and 3 negative (×) per cell for CTC. In contrast, BoxTrack requires only a single bounding box per cell in the first frame. Despite this drastic reduction in manual effort, BoxTrack achieves +**15.96** MOTA improvement on CTMC and +**8.86** on CTC over SAT.

One such method is SAT (Segment and Track Anything) [10], which proposes an efficient pipeline using sparse point-based supervision in only the first frame. Specifically, SAT requires annotating both positive points (inside each cell) and

negative points (background) to guide segmentation and tracking. Despite its annotation efficiency compared to fully supervised pipelines, SAT still involves considerable manual effort, particularly for dense or large-scale datasets where each cell demands multiple annotated points.

This work introduces **BoxTrack**, a simplified yet highly effective tracking framework that eliminates the need for point-based supervision. BoxTrack relies solely on bounding box annotations in the first frame of each sequence—one box per cell—and leverages modern segmentation and tracking mechanisms to propagate this minimal supervision across the full sequence. This approach significantly reduces annotation overhead, making it highly scalable and well-suited for large microscopy datasets.

Fig. 1 compares the annotation burden of SAT and BoxTrack on CTMC [1] and CTC [28]. SAT requires 30 positive (•) and 3 negative (×) points per cell for CTMC, and 6 positive plus 3 negative points for CTC. BoxTrack, in contrast, uses only one bounding box per cell in the first frame. Despite this simplification, it achieves a MOTA improvement of **15.96%** on CTMC and **8.86%** on CTC, highlighting box-level supervision as an efficient and scalable alternative for cell tracking in microscopy. Removing the need for point-level labels enables broader applicability in biomedical workflows and large-scale screening. The main contributions of this work are as follows:

- **BoxTrack**: a weakly supervised cell tracking framework using only one bounding box per cell in the first frame.
- Achieves over 6× annotation savings compared to SAT [10], with improved tracking accuracy.
- Outperforms SAT with MOTA gains of **15.96%** (CTMC) and **8.86%** (CTC).
- Generalizes across imaging modalities and cell types (DeepCell results).
- Maintains strong performance under reduced scan frequency, supporting long-term imaging up to 60-minute intervals.

The remainder of the paper is organized as follows: Section 2 reviews related work with a focus on weakly supervised cell tracking. Section 3 details the BoxTrack framework. Section 4 describes the evaluation datasets (CTMC, CTC, DeepCell), and Section 5 outlines the experimental setup. Section 6 reports the results and analysis. Section 7 concludes with key takeaways and future directions for scalable tracking.

2 Related Work

2.1 Cell Tracking and Segmentation

Cell tracking is vital for analyzing proliferation, migration, and interactions [5, 28]. Traditional segmentation-first methods [8, 11–13, 24, 26] require dense labels and struggle to generalize without retraining. Recent models like Trackastra [6], CellTrack R-CNN [3], and SC-Track [17] integrate tracking and segmentation, improving lineage accuracy but still demand high supervision. This highlights the need for scalable, annotation-efficient tracking approaches.

2.2 Weakly Supervised Approaches

Weakly supervised methods have emerged to reduce the annotation burden while preserving accuracy. Point-supervised models [9,14] and box-based techniques [7] offer lighter alternatives to full masks. However, they often depend on pre-trained backbones or complex post-processing, limiting adaptability. Segment and Track Anything (SAT) [10] reduces the labeling per frame by using positive and negative points in the first frame to track cells over time. While efficient, SAT still requires around 30 clicks per cell and can struggle in crowded or noisy settings, where point placement and boundary precision become critical.

2.3 Challenges in Microscopy

Microscopy data presents unique challenges not seen in natural images, including low contrast, diverse modalities (e.g., fluorescence, phase contrast), and densely packed or overlapping cells [25, 30, 32]. General-purpose models such as YOLO [29] and SAM [15] require significant adaptation to perform reliably in biomedical settings. In addition, long sequences with cell division or morphological changes complicate the tracking. Although domain-specific models such as Trackastra [6] improve performance through customized designs, they still rely on dense supervision and often lack generalizability across modalities.

2.4 Need for BoxTrack

While weak supervision has reduced annotation costs, existing approaches often trade off accuracy or require dense point-level input, which becomes impractical in crowded cell environments. To address this, we introduce *BoxTrack*—a simple yet effective tracking framework that requires only a single bounding box per cell in the first frame. It avoids reliance on point annotations, segmentation masks, or modality-specific tuning, instead leveraging a unified detection-based strategy to track cells across time.
As shown in Fig. 1, BoxTrack drastically lowers annotation effort while outperforming SAT on CTMC and CTC benchmarks, demonstrating that minimal supervision can achieve both accuracy and cross-modal generalizability.

3 BoxTrack: The Proposed Approach

BoxTrack is a zero-shot, inference-only framework for cell segmentation and tracking in microscopy sequences. It requires a single bounding-box prompt per cell in the first frame and generalizes effectively across diverse cell cultures and imaging conditions without retraining. BoxTrack consists of two main modules, as visualized in Fig. 2: the **Segmentation Module** (purple), responsible for generating per-frame mask proposals, and the **Tracking Module** (green), which links these proposals across frames to form consistent trajectories.

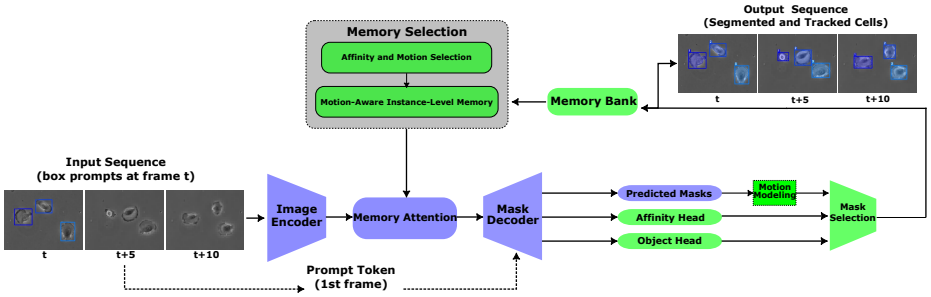

Fig. 2. Overview of the BoxTrack pipeline. Given a video and bounding-box prompts per cell in the first frame only, the **Segmentation Module (purple)** uses a frozen SAM2 backbone [22] to produce object tokens and mask proposals. These feed into the **Tracking Module (green)**, where Motion Modeling with a Kalman filter predicts spatial movement, and a memory-aware scoring mechanism selects the final ID-annotated masks and trajectories [31].

3.1 Segmentation Module

The Segmentation Module leverages a frozen backbone derived from SAM2 [22], pretrained on the extensive SA-V dataset. For architectural details of SAM2, including its Image Encoder and prompt fusion design, we refer the reader to the original paper. This module produces object tokens, intermediate masks, and bounding boxes in each frame. As shown in Fig. 2, it comprises the **Prompt Encoder**, **Image Encoder**, **Memory Attention** block, and **Mask Decoder**, which operate sequentially to extract and refine segmentation features.
In the first frame I_1 of a microscopy sequence $V = \{I_1, I_2, \ldots, I_T\}$, bounding boxes $\{b_i^1\}_{i=1}^N$ are annotated to initialize tracking. These provide explicit spatial cues for each cell. A lightweight **Prompt Encoder** converts the bounding boxes into sparse object-specific embeddings $\mathbf{P}_0$:

$$\mathbf{P}_0 = \text{PromptEncoder}(b_1^1, \ldots, b_N^1). \tag{1}$$

These embeddings encapsulate positional and visual context and guide the segmentation process.
Each frame I_t is processed by the frozen **Image Encoder** to extract spatial visual features:

$$\mathbf{F}_t = \text{ImageEncoder}(I_t), \quad \mathbf{F}_t \in \mathbb{R}^{H \times W \times C}, \tag{2}$$

where H and W denote spatial resolution and C the number of feature channels. The extracted features $\mathbf{F}_t$ are fused with prompt embeddings $\mathbf{P}_0$ and memory tokens from the previous frame, selected by the Tracking Module via the **Memory Attention** block. This fusion incorporates temporal context for refinement. The result is passed to the **Mask Decoder** to generate object tokens $\mathbf{z}_i^t$ and

corresponding segmentation masks $\hat{M}_i^t$:

$$\hat{M}_i^t, \mathbf{z}_i^t = \text{MaskDecoder}(\mathbf{P}_0, \mathbf{F}_t, \text{MemoryAttention}). \tag{3}$$

Bounding boxes $\hat{b}_i^t$ are derived from these masks and forwarded to the tracking stage.

3.2 Tracking Module

The Tracking Module links segmentation proposals across frames to ensure temporal consistency, especially in the presence of occlusions and densely packed cells [31]. It consists of two main components: a **Motion Modeling** block, implemented using a **Kalman Filter**, and a **Memory-Aware Mask Selection** block. The **Motion Modeling** block predicts the expected position of each cell in the next frame using the bounding box $\hat{b}_i^{t-1}$ from the previous frame:

$$\tilde{b}_i^t = \text{KalmanPredict}(\hat{b}_i^{t-1}). \tag{4}$$

This prediction serves as a motion prior to reduce association ambiguity by constraining the search space.

Unlike SAM2 [22], which maintains memory internally within the segmentation stream using a fixed-size window (typically 7 frames) and a First-In, First-Out (FIFO) policy that discards the oldest frame regardless of quality, BoxTrack separates memory handling into the Tracking Module. This decoupling enables more flexible and robust memory management. Specifically, when a high-quality candidate is identified, it is added to the memory bank if space is available. If the memory is full, the stored frame with the lowest combined confidence score—recorded at the time of insertion—is evicted and replaced. This score-based replacement strategy actively curates memory content and improves resilience to appearance changes and occlusions, offering a key advantage over FIFO-based approaches.

The memory tokens selected by this process are fed into the Segmentation Module's **Memory Attention** block, enriching it with temporally relevant information for the next frame's prediction.

To determine which mask proposals to retain, the **Memory-Aware Mask Selection** block computes a confidence score s_i^t for each candidate by combining three components:

the **Affinity Score**, measuring token similarity across frames:

$$s_{a,i}^t = \cos(\mathbf{z}_i^t, \mathbf{z}_i^{t-1}), \tag{5}$$

the **Motion Score**, evaluating mask overlap between consecutive frames:

$$s_{m,i}^t = \text{IoU}(\hat{M}_i^t, \hat{M}_i^{t-1}), \tag{6}$$

and the **Objectness Score**, estimated from the current object token via a small multi-layer perceptron (MLP) followed by sigmoid activation:

$$s_{o,i}^t = \sigma(\text{MLP}(\mathbf{z}_i^t)). \tag{7}$$

These are aggregated into a final score:

$$s_i^t = \lambda_a s_{a,i}^t + \lambda_m s_{m,i}^t + \lambda_o s_{o,i}^t, \quad \text{where } \lambda_a + \lambda_m + \lambda_o = 1. \tag{8}$$

The proposal with the highest s_i^t is retained, and its object token is stored in memory for use in the next frame. For each time step, the selected segmentation masks $\hat{M}_i^t$ are assigned persistent track IDs, yielding the **final output** of BoxTrack: a temporally linked sequence of instance segmentation masks across the video.

4 Datasets

BoxTrack builds on a segmentation backbone derived from SAM2 [22], trained on the SA-V dataset comprising 50.9K videos and 35.5M instance masks across 642K masklets.
Evaluation is performed on three microscopy benchmarks. The CTMC dataset [1] includes 86 videos from 14 cell lines; 22 sequences are selected to represent diverse imaging conditions. The CTC dataset [18] provides 2D/3D time-lapse sequences acquired using Bright Field, Phase Contrast, and DIC microscopy. Four 2D sequences are used, totaling 8,017 frames with an average of 33.12 cells per frame. The DeepCell dataset [21] offers 12 test sequences comprising 617 frames and 99,550 annotated cells, enabling evaluation across dense nuclear environments.
In contrast to prior SAM-based methods such as SAT [10], which require supervised training on microscopy data (e.g., LIVECell [5]), BoxTrack operates in a training-free manner using only a single bounding box per cell—demonstrating high generalization with minimal supervision.

5 EXPERIMENTAL SETUP

BoxTrack is evaluated under four experimental settings. The *Wide-Ranging Cell Types (CTMC)* setting uses 22 sequences from the CTMC dataset [1] to benchmark BoxTrack against SAT across diverse cell morphologies and motion behaviors. BoxTrack achieves higher tracking accuracy while requiring significantly fewer annotations. The *Multi-Modality Imaging (CTC)* setting includes 2D sequences from the CTC dataset [18], spanning Phase Contrast, Bright Field, and Fluorescent modalities. Here too, BoxTrack consistently outperforms SAT under varying imaging conditions. The *Fluorescent Nuclear Tracking (DeepCell)* setting involves 12 sequences (617 frames, 99,550 annotated cells) from the DeepCell test set [21], testing generalization across acquisition protocols and nuclear appearances. The *Temporal Downsampling* setting subsamples CTMC sequences at intervals up to 60 minutes to evaluate performance with sparse temporal input. Unlike SAT, which is fine-tuned on LIVECell [5], BoxTrack operates without retraining and uses only a single bounding box per cell in the first frame. Performance is evaluated using standard MOT metrics [2, 16, 23], including MOTA (Multi-Object Tracking Accuracy), IDF1 (ID-based F1 score),

IDS (identity switches), MT (mostly tracked), and ML (mostly lost). For MOTA and related formulas, please refer to [2]. For mathematical definitions, including the MOTA formula, we refer readers to the original CLEAR MOT paper [2]. All experiments were carried out with Python 3.10, PyTorch 2.3.1, and TorchVision 0.18.1. The segmentation backbone (SAM2.1-hiera-large) is used in inference-only mode.

5.1 Multi-Cell Type Tracking (CTMC)

The CTMC dataset [1] comprises 22 sequences from diverse cell lines exhibiting diverse morphologies, densities, and motion patterns. This setting assesses generalization across varied biological conditions. As shown in Table 1, BoxTrack consistently outperforms SAT in both MOTA and IDF1, with substantial gains observed in sequences like `BPAE-run05` (+56.13% MOTA) and `A-10-run05` (+19.74% MOTA). While SAT occasionally reports higher IDF1 (e.g., `CRE-BAG2-run03`), BoxTrack demonstrates more stable overall tracking performance.

Table 1. Comparison of BoxTrack and SAT across CTMC dataset. Bold indicates better performance per metric. BoxTrack uses bounding box annotations denoted by ß, while SAT uses point supervision denoted by $P(X)$, where X is the number of annotated points per cell.

Sequence	**Supervision**		**MOTA (%) ↑**		**IDF1 (%) ↑**		**IDS ↓**		**MT (%) ↑**		**ML (%) ↓**	
	BoxTrack	SAT	BoxTrack	SAT	BoxTrack	SAT	BoxTrack	SAT	BoxTrack	SAT	BoxTrack	SAT
3T3-run03	ß	$P(18)$	**93.01**	61.49	**88.72**	87.50	2.00	**0.00**	100.00	100.00	0.00	0.00
A-10-run01	ß	$P(23)$	**83.67**	80.78	**91.82**	90.56	0.00	0.00	80.00	80.00	0.00	0.00
A-10-run05	ß	$P(33)$	**99.70**	79.96	**99.85**	95.56	0.00	0.00	100.00	100.00	0.00	0.00
A-10-run07	ß	$P(23)$	**93.87**	77.81	**96.94**	91.25	0.00	0.00	85.71	85.71	0.00	0.00
A-549-run03	ß	$P(28)$	**89.13**	57.51	79.39	**82.00**	2.00	**0.00**	88.89	88.89	0.00	0.00
BPAE-run05	ß	$P(33)$	**82.50**	26.37	**81.16**	77.14	3.00	**0.00**	100.00	100.00	0.00	0.00
CRE-BAG2-run03	ß	$P(23)$	**46.22**	36.62	62.45	**72.41**	1.00	**0.00**	60.00	60.00	6.67	6.67
LLC-MK2-run01	ß	$P(33)$	**76.49**	54.28	68.67	**80.65**	1.00	**0.00**	75.00	75.00	0.00	0.00
LLC-MK2-run02a	ß	$P(33)$	**92.61**	60.06	**95.46**	83.87	1.00	**0.00**	100.00	100.00	0.00	0.00
LLC-MK2-run03	ß	$P(28)$	**96.40**	96.18	**98.20**	98.11	0.00	0.00	100.00	100.00	0.00	0.00
APM-run05	ß	$P(28)$	43.79	**61.85**	69.59	**82.38**	1.00	**0.00**	50.00	**75.00**	0.00	0.00
LLC-MK2-run07	ß	$P(18)$	**77.32**	73.72	**88.66**	88.65	0.00	0.00	77.78	77.78	0.00	0.00
MDBK-run03	ß	$P(28)$	**87.11**	69.76	**93.68**	86.00	0.00	0.00	82.35	82.35	0.00	0.00
MDBK-run09	ß	$P(23)$	**79.77**	53.28	**89.89**	85.88	2.00	**0.00**	80.00	80.00	0.00	0.00
MDOK-run07	ß	$P(18)$	**100.00**	99.80	**100.00**	99.80	0.00	0.00	100.00	100.00	0.00	0.00
OK-run01	ß	$P(18)$	**81.52**	60.94	**90.68**	85.39	0.00	0.00	80.00	80.00	0.00	0.00
OK-run05	ß	$P(33)$	**87.11**	65.11	**93.56**	91.30	0.00	0.00	100.00	100.00	0.00	0.00
OK-run07	ß	$P(18)$	**70.78**	19.52	**85.39**	72.22	0.00	0.00	68.75	68.75	0.00	0.00
PL1Ut-run05	ß	$P(33)$	**99.46**	93.13	**99.73**	96.55	0.00	0.00	100.00	100.00	0.00	0.00
RK-13-run03	ß	$P(33)$	**85.83**	65.11	85.67	85.67	2.00	**0.00**	85.71	85.71	0.00	0.00
U2O-S-run03	ß	$P(28)$	**81.55**	76.58	**90.77**	85.00	0.00	0.00	75.00	75.00	0.00	0.00
U2O-S-run05	ß	$P(33)$	**57.74**	54.44	78.87	**84.62**	0.00	0.00	62.50	62.50	0.00	0.00
Average	–	–	**82.79**	66.83	**87.91**	86.36	0.61	**0.00**	86.28	**86.79**	0.29	0.29

5.2 Multi-Modality Imaging (CTC)

This setting uses 2D sequences from the CTC dataset [18], spanning Phase Contrast, Bright Field, and Fluorescent microscopy. It evaluates robustness across

distinct imaging modalities. Table 2 shows that BoxTrack outperforms SAT across all sequences in MOTA and IDF1. Notable gains are seen in `PhC-C2DH-U373 (01)` and `Fluo-N2DH-GOWT1`, with reduced Mostly Lost rates further indicating improved temporal consistency under modality shifts.

Table 2. Comparison of BoxTrack and SAT across CTC dataset. Bold indicates better performance per metric. BoxTrack uses bounding box annotations denoted by ß, while SAT uses point supervision denoted by $P(X)$, where X is the number of annotated points per cell.

Sequence	**Supervision**		**MOTA (%) ↑**		**IDF1 (%) ↑**		**IDS ↓**		**MT (%) ↑**		**ML (%) ↓**	
	BoxTrack	SAT	BoxTrack	SAT	BoxTrack	SAT	BoxTrack	SAT	BoxTrack	SAT	BoxTrack	SAT
PhC-C2DH-U373 (01)	ß	$P(9)$	**100.00**	79.82	**100.00**	89.00	0.00	0.00	**100.00**	71.43	0.00	0.00
PhC-C2DH-U373 (02)	ß	$P(9)$	**85.71**	82.75	**93.33**	91.94	0.00	0.00	100.00	100.00	0.00	0.00
Fluo-N2DH-GOWT1	ß	$P(6)$	**97.81**	88.24	**98.91**	91.20	0.00	0.00	**91.67**	79.17	**0.00**	8.34
Fluo-N2DH-SIM+	ß	$P(6)$	**85.79**	83.03	**90.44**	88.73	0.00	0.00	**83.84**	83.34	0.00	0.00
Average	–	–	**92.32**	83.46	**95.67**	90.22	**0.00**	0.00	**93.87**	83.45	0.00	2.08

Table 3. Overall performance of BoxTrack-SAT on the DeepCell tracking dataset.

Method	**MOTA (%) ↑**	**IDF1 (%) ↑**	**IDS ↓**	**MT (%) ↑**	**ML (%) ↓**
BoxTrack-SAT	82.95	87.17	15.67	68.67	19.13

5.3 Multi-Culture Nuclear Imaging (DeepCell)

The DeepCell DynamicNuclearNet dataset [21] contains fluorescent nuclear images from multiple mammalian cell cultures, introducing variation in nuclear density, shape, and arrangement. This setting tests generalization to unseen visual and biological characteristics. BoxTrack achieves a MOTA of 82.95 and IDF1 of 87.17 without retraining (Table 3), highlighting its robustness in dense nuclear tracking tasks.

5.4 Temporal Resolution Sensitivity (CTMC-Frequency)

This setting evaluates the impact of scan interval length using sub-sampled CTMC sequences. The original acquisition rate of one frame every 30 seconds is progressively reduced up to one frame every 60 minutes. As shown in Fig. 3, MOTA gradually declines from 72.01% at 1-minute intervals to 55.93% at 60 minutes. Despite the drop in temporal density, BoxTrack maintains strong performance, demonstrating resilience in low-frequency or long-term imaging conditions.

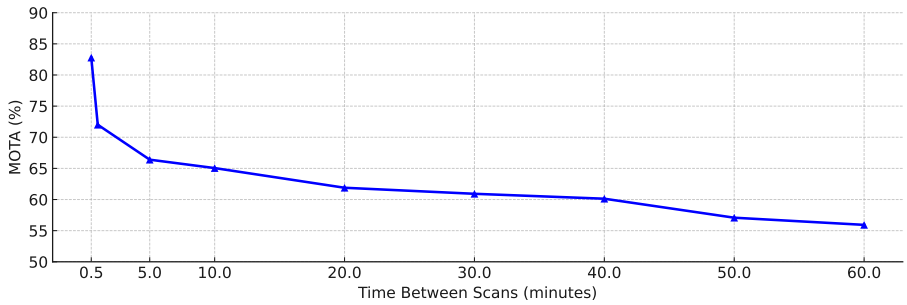

Fig. 3. Tracking performance (MOTA) of BoxTrack across different temporal intervals on CTMC.

6 Analysis and Discussion

This section evaluates the effectiveness of BoxTrack across four experimental settings, each reflecting distinct challenges in microscopy-based cell tracking.

6.1 CTMC: Cell Line Diversity and Tracking Generalization

The Multi-Cell Type Tracking setting used 22 sequences from diverse cell lines in the CTMC dataset. BoxTrack consistently outperformed SAT with an average MOTA improvement of +15.96%. For example, in the `OK-run07` sequence (shown in Fig. 4, top), BoxTrack achieved a MOTA of 70.78%, significantly higher than SAT's 19.52. At $t+790$ minutes, SAT failed to detect and track cell ID 10, which was correctly identified by BoxTrack. By $t + 975$, SAT missed cells 10, 6, 11, 1, and 5—all of which were successfully tracked by BoxTrack. Additional missed detections and less precise boundaries were observed in SAT.
These results highlight BoxTrack's scalability across biologically diverse conditions without the need for retraining or dataset-specific adaptation. Such capabilities are particularly useful in large-scale experiments and high-throughput time-lapse studies, where annotation and tuning overheads are often prohibitive.

6.2 CTC: Robustness Across Imaging Modalities

The *Multi-Modality Imaging* setting assessed performance across Phase Contrast, Bright Field, and Fluorescent microscopy using sequences from the CTC dataset. BoxTrack maintained modality-agnostic performance, achieving average MOTA and IDF1 scores of 92.32% and 95.67% respectively, compared to SAT's 83.46 and 90.22. In the `PhC-C2DH-U373` sequence (Fig. 4, bottom), SAT achieved a MOTA of 79.82%, whereas BoxTrack reached 100.00%. At $t + 495$ minutes, SAT inaccurately delineated the boundary of cell ID 4, and this error persisted through $t+690$. In contrast, BoxTrack preserved precise and consistent tracking

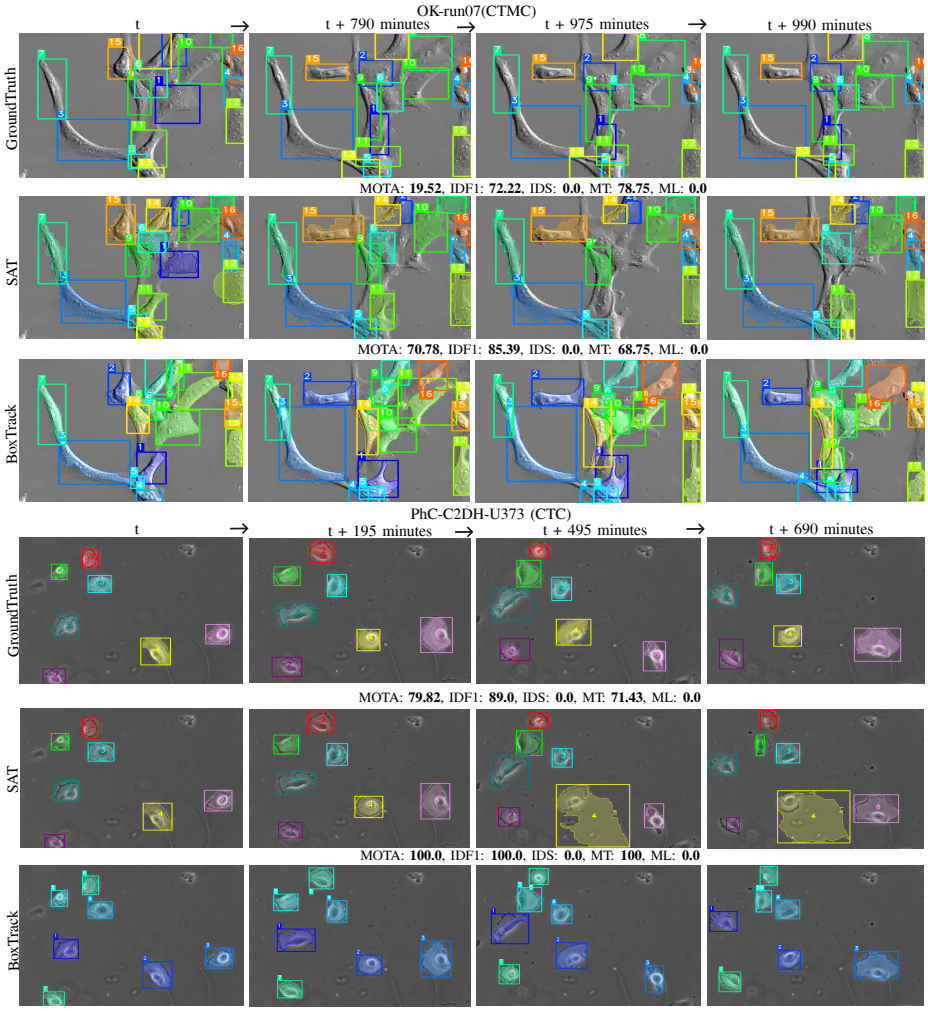

Fig. 4. Qualitative comparison for the first two experimental settings: **CTMC (Top)** and **CTC (Bottom)**. Each row shows Ground Truth, SAT, and BoxTrack predictions over time. Tracking metrics are listed above SAT and BoxTrack rows. BoxTrack consistently outperforms SAT across diverse cell types and imaging modalities.

throughout the sequence.

The results underline the robustness of BoxTrack under modality shifts and suggest that the approach can be applied across varying acquisition setups without handcrafted modifications. This simplifies its use in multi-institutional studies or labs employing diverse imaging technologies.

6.3 DeepCell: Scalability in Dense Nuclear Tracking

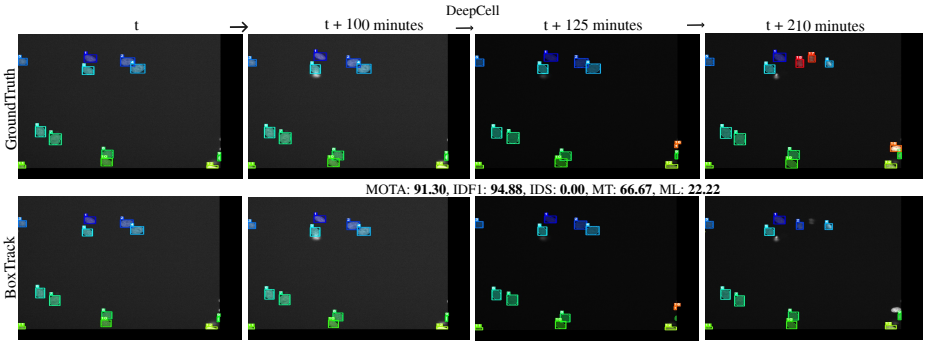

Fig. 5. Qualitative comparison for the third experimental setting: **DeepCell**. Each row shows Ground Truth and BoxTrack predictions over time. Tracking metrics are reported above BoxTrack row. BoxTrack shows strong performance in densely populated, morphologically varied nuclear cultures.

The *Multi-Culture Nuclear Imaging* setting focused on nuclear tracking under fluorescence microscopy across multiple cell cultures. Despite consistent modality, variations in nuclear shape, density, and distribution introduced additional complexity. BoxTrack achieved strong performance with a MOTA of 82.95% and IDF1 of 87.17% across 12 sequences. As shown in Fig. 5, BoxTrack maintained robust accuracy despite contrast changes between $t + 100$ and $t + 125$ minutes. However, at $t + 210$ minutes, the model failed to detect a mitotic event where cell ID 2 divided into cells 17 and 18.
These findings suggest that BoxTrack is well-suited for large-scale nuclear studies in areas such as drug screening or cancer biology, where high-density and morphologically varied samples are common. Future integration of division-aware tracking could enhance its applicability in proliferative assays.

6.4 Temporal Downsampling: Robustness to Reduced Scan Frequency

The *Temporal Resolution Sensitivity* setting tested BoxTrack under lower scan frequencies by subsampling CTMC sequences from 30 seconds up to 60-minute intervals. As expected, MOTA gradually declined with reduced temporal resolution—from 82.79% at 30 seconds to 55.93% at 60 minutes. Despite fewer temporal cues, BoxTrack preserved a MOTA of 60.92% even at 30-minute intervals and remained above 55% at 60 minutes.
These results indicate that BoxTrack remains effective in temporally sparse conditions, making it suitable for long-term imaging studies where frequent scanning

is impractical due to phototoxicity, memory constraints, or biological limitations. Across all experiments, BoxTrack consistently delivered strong tracking with minimal supervision—only one bounding box per cell—unlike SAT, which requires dense points and backbone tuning. Its zero-shot capability, scalability, and low annotation effort make BoxTrack well-suited for real-world scenarios, including live-cell imaging, large datasets, varied conditions, and low-frequency acquisitions.

7 Conclusion

This work introduced BoxTrack, a training-free, annotation-efficient framework for cell tracking in microscopy images that relies solely on a single bounding box per cell in the first frame. Through extensive experiments across three diverse settings—cell line diversity (CTMC), modality variation (CTC), and nuclear culture dynamics (DeepCell)—BoxTrack consistently outperformed the performance of SAT, a recent state-of-the-art weakly supervised baseline. The results demonstrate that BoxTrack not only reduces manual annotation time by a substantial margin but also generalizes robustly across different cell types, densities, and imaging conditions without retraining or fine-tuning. This highlights the potential of BoxTrack as a universal tracking pipeline for microscopy, bridging the gap between usability and performance. By offering a lightweight yet accurate alternative to heavily supervised or fine-tuned systems, BoxTrack paves the way for broader accessibility in live-cell imaging studies, high-throughput screens, and biomedical applications where resource constraints or annotation budgets are critical bottlenecks. Future work may further explore integration with cell lineage tracing, event detection (e.g., mitosis), and real-time deployment in automated microscopy workflows.

Acknowledgements

This work is partially funded by SAIL (Sartorius AI Lab), a collaboration between the German Research Center for Artificial Intelligence (DFKI) and Sartorius AG.

References

1. Anjum, S., Gurari, D.: Ctmc: Cell tracking with mitosis detection dataset challenge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 982–983 (2020)
2. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing (2008)
3. Chen, Y., Song, Y., Zhang, C., Zhang, F., O'Donnell, L., Chrzanowski, W., Cai, W.: Celltrack r-cnn: A novel end-to-end deep neural network for cell segmentation and tracking in microscopy images. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 779–782. IEEE (2021)

4. Cheng, B., Parkhi, O., Kirillov, A.: Pointly-supervised instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2617–2626 (2022)
5. Edlund, C., Jackson, T.R., Khalid, N., Bevan, N., Dale, T., Dengel, A., Ahmed, S., Trygg, J., Sjögren, R.: Livecell—a large-scale dataset for label-free live cell segmentation. Nature methods (2021)
6. Gallusser, B., Weigert, M.: Trackastra: Transformer-based cell tracking for live-cell microscopy. In: European Conference on Computer Vision. pp. 467–484. Springer (2024)
7. Khalid, N., Caroprese, M., Lovell, G., Porto, D.A., Trygg, J., Dengel, A., Ahmed, S.: Bounding box is all you need: Learning to segment cells in 2d microscopic images via box annotations. In: Annual Conference on Medical Image Understanding and Analysis. pp. 314–328. Springer (2024)
8. Khalid, N., Caroprese, M., Lovell, G., Trygg, J., Dengel, A., Ahmed, S.: Cellspot: Deep learning-based efficient cell center detection in microscopic images. In: International Conference on Artificial Neural Networks. pp. 215–229. Springer (2024)
9. Khalid, N., Froes, T.C., Caroprese, M., Lovell, G., Trygg, J., Dengel, A., Ahmed, S.: Pace: Point annotation-based cell segmentation for efficient microscopic image analysis. In: International Conference on Artificial Neural Networks. Springer (2023)
10. Khalid, N., Koochali, M., Naseem, K., Caroprese, M., Lovell, G., Porto, D.A., Trygg, J., Dengel, A., Ahmed, S.: Sat: Segment and track anything for microscopy. In: Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART. pp. 286–297. INSTICC, SciTePress (2025). https://doi.org/10.5220/0013154200003890
11. Khalid, N., Koochali, M., Rajashekar, V., Munir, M., Edlund, C., Jackson, T.R., Trygg, J., Sjögren, R., Dengel, A., Ahmed, S.: Deepmucs: A framework for coculture microscopic image analysis: From generation to segmentation. In: 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE (2022)
12. Khalid, N., Munir, M., Edlund, C., Jackson, T.R., Trygg, J., Sjögren, R., Dengel, A., Ahmed, S.: Deepcens: An end-to-end pipeline for cell and nucleus segmentation in microscopic images. In: 2021 International Joint Conference on Neural Networks (IJCNN). IEEE (2021)
13. Khalid, N., Munir, M., Edlund, C., Jackson, T.R., Trygg, J., Sjögren, R., Dengel, A., Ahmed, S.: Deepcis: An end-to-end pipeline for cell-type aware instance segmentation in microscopic images. In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE (2021)
14. Khalid, N., Schmeisser, F., Koochali, M., Munir, M., Edlund, C., Jackson, T.R., Trygg, J., Sjögren, R., Dengel, A., Ahmed, S.: Point2mask: A weakly supervised approach for cell segmentation using point annotation. In: Medical Image Understanding and Analysis: 26th Annual Conference, MIUA 2022, Cambridge, UK, July 27–29, 2022, Proceedings. Springer (2022)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
16. Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.: Motchallenge 2015: Towards a benchmark for multi-target tracking. arXiv preprint arXiv:1504.01942 (2015)

17. Li, C., Xie, S.S., Wang, J., Sharvia, S., Chan, K.Y.: Sc-track: a robust cell-tracking algorithm for generating accurate single-cell lineages from diverse cell segmentations. Briefings in Bioinformatics **25**(3), bbae192 (2024)
18. Maška, M., Ulman, V., Delgado-Rodriguez, P., Gómez-de Mariscal, E., Nečasová, T., Guerrero Peña, F.A., Ren, T.I., Meyerowitz, E.M., Scherr, T., Löffler, K., et al.: The cell tracking challenge: 10 years of objective benchmarking. Nature Methods **20**(7), 1010–1020 (2023)
19. Masuzzo, P., Van Troys, M., Ampe, C., Martens, L.: Taking aim at moving targets in computational cell migration. Trends in cell biology **26**(2), 88–110 (2016)
20. Meijering, E., Dzyubachyk, O., Smal, I., van Cappellen, W.A.: Tracking in cell and developmental biology. In: Seminars in cell & developmental biology. vol. 20, pp. 894–902. Elsevier (2009)
21. Moen, E., Borba, E., Miller, G., Schwartz, M., Bannon, D., Koe, N., Camplisson, I., Kyme, D., Pavelchek, C., Price, T., et al.: Accurate cell tracking and lineage construction in live-cell imaging experiments with deep learning. Biorxiv p. 803205 (2019)
22. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
23. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision. Springer (2016)
24. Scherr, T., et al.: Cell segmentation and tracking using cnn-based distance predictions and a graph-based matching strategy. PloS one **15**(12), e0243219 (2020)
25. Stringer, C., Wang, T., Michaelos, M., Pachitariu, M.: Cellpose: a generalist algorithm for cellular segmentation. Nature Methods (2020)
26. Stringer, C., et al.: Cellpose: a generalist algorithm for cellular segmentation. Nature Methods **18**(1), 100–106 (2021)
27. Tian, Z., Shen, C., Wang, X., Chen, H.: Boxinst: High-performance instance segmentation with box annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5443–5452 (2021)
28. Ulman, V., Maška, M., Magnusson, K.E., Ronneberger, O., Haubold, C., Harder, N., Matula, P., Matula, P., Svoboda, D., Radojevic, M., et al.: An objective comparison of cell-tracking algorithms. Nature methods (2017)
29. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7464–7475 (2023)
30. Wang, R., Butt, D., Cross, S., Verkade, P., Achim, A.: Bright-field to fluorescence microscopy image translation for cell nuclei health quantification. Biological Imaging **3**, e12 (2023)
31. Yang, C.Y., Huang, H.W., Chai, W., Jiang, Z., Hwang, J.N.: Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. arXiv preprint arXiv:2411.11922 (2024)
32. Yazdi, R., Khotanlou, H.: A survey on automated cell tracking: challenges and solutions. Multimedia Tools and Applications (2024)