CytoTracker: Multi-Cell Tracking with Diffusion-Guided Cell Association

Komal Kumar[1] KOMAL.KUMAR@MBZUAI.AC.AE
[1] *Mohamed Bin Zayed University of Artificial Intelligence, UAE*

Noa Novershtern[2]
[2] *Weizmann Institute of Science, Israel* NOA.NOVERSHTERN@WEIZMANN.AC.IL
Jacob Hanna[2] JACOB.HANNA@WEIZMANN.AC.IL
[2] *Weizmann Institute of Science, Israel*

Hisham Cholakkal[1] HISHAM.CHOLAKKAL@MBZUAI.AC.AE
[1] *Mohamed Bin Zayed University of Artificial Intelligence, UAE*

Editors: Under Review for MIDL 2026

Abstract

Cell tracking is a challenging task in microscopic image analysis that enables the study of cellular behaviors and interactions over time. It requires tracking hundreds of nearly indistinguishable cells that move dynamically and may undergo growth or division, complicating the tracking process. As a result, state-of-the-art cell tracking methods often rely on ground truth segmentation mask-derived cell-level features, even during inference, which are typically unavailable in real-world application. In this work, we introduce CytoTracker, an end-to-end framework for cell tracking that integrates three key modules: a cell detector, a diffusion model-based motion prediction network, and a transformer-based association network. Additionally, we developed the CytoEmbedding attention fusion block to effectively extract cell-level features from images, improving tracking. Our method accurately detects cells in each frame, predicts nonlinear motion, and robustly associates them across time, even in the presence of cell division. Experimental results on microscopy datasets demonstrate that CytoTracker achieves performance comparable to state-of-the-art approaches that require ground truth masks during inference, without such costly segmentation mask annotations.

Keywords: cell detection, lineage tracking, cell association, diffusion motion.

1. Introduction

Cell tracking is a challenging medical image analysis task, essential for studying cellular behaviours, dynamics, and interactions over time. It plays a key role in diverse applications ranging from developmental biology to cancer research. In cell tracking, the objective is to track individual cell instances across frames in a microscopic video, where these cell instances need to be associated across frames despite occurrences of cell division, migration, and apoptosis. Cell tracking presents unique challenges due to the large number of similar-looking cells exhibiting complex behaviours, including cell division and merging (see Fig. 1). These challenges make developing robust and accurate cell tracking methods particularly demanding. Similar to cell tracking in microscopy, multiple object tracking (MOT) in natural images also aims to track multiple objects in a video. Existing MOT methods for natural images have incorporated deep learning-based motion models, which can be

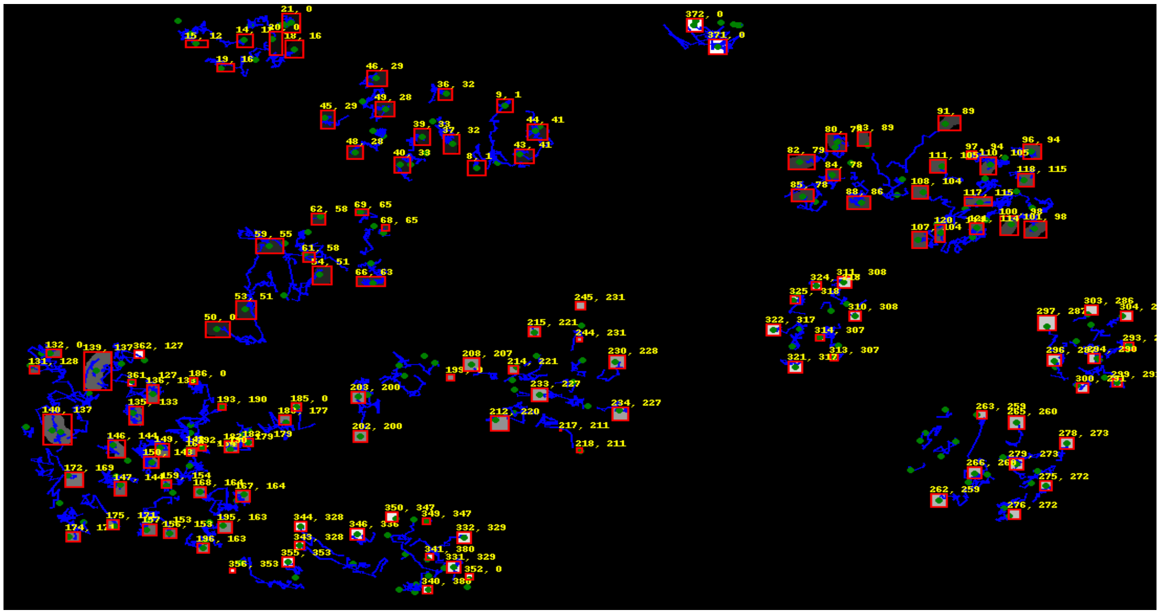

Figure 1: Visualization of cell detection and tracking trajectories. Each cell is represented with a unique ID, and its movement across frames is depicted by blue trajectory lines. Red bounding boxes indicate detected cells, while green dots mark their current positions. A notation such as 132, 0 represents a cell that has not undergone division, while 362, 127 indicates a cell that originated from parent cell 127 (Best viewed in Zoom).

broadly categorized into linear and non-linear motion estimation techniques. Linear models, such as SORT (Bewley et al., 2016), DeepSORT (Wojke et al., 2017), ByteTrack (Zhang et al., 2022), and OC-SORT (Cao et al., 2023), assume smooth motion within small time intervals. Although these models achieve high-speed tracking, they lack the flexibility to handle unpredictable cellular movements. On the other hand, non-linear motion prediction techniques, such as tracking optical flow (Xiao et al., 2024), DiffTrack (Lv et al., 2024), and transformer models (Cao et al., 2022; Gallusser and Weigert, 2024), can effectively model complex motion dynamics in natural images. However, performing cell association through these non-linear methods also remains a challenge due to frequent occlusions and the presence of hundreds of cell divisions, making it difficult to accurately link new cells to their parents. These challenges make adapting existing MOT methods for the cell tracking problem particularly difficult.

Several cell tracking approaches have been proposed in the literature to address the challenges related to the tracking/association of cell instances across video frames. For example, integer programming-based lineage tracing (Magnusson et al., 2015) and graph-based models (Ben-Haim and Raviv, 2022) have been developed by extracting handcrafted features and manually tuning them. Similarly, the robustness of cell tracking was improved in (Chen et al., 2020), while transformer-based tracking was leveraged in (Sun et al., 2020) to effectively handle occlusions and long-term dependencies. However, these methods often struggle to associate dividing cells (Zhang and Yang, 2023). Graph Neural Networks (GNNs) (Moen et al., 2019; Ben-Haim and Raviv, 2022) refine object interactions but are restricted to local regions, unlike transformers, which enable all-to-all associations. Methods such as (O'Connor et al., 2022) predict object masks for linking. Recently, the state-of-the-art method Trackastra (Gallusser and Weigert, 2024) learns associations using cell-level features extracted from *manually labeled ground truth segmentation masks*. However, these *ground truth masks are required by this model, even during inference*.

In this work, we propose CytoTracker, which strives to address the limitations of existing cell tracking approaches, through an end-to-end cell tracking framework that integrates detection, motion prediction, and association within a unified architecture. Specifically, our CytoTracker employs YOLOX (Ge et al., 2021) for accurate frame-wise cell detection and then utilizes a diffusion-based motion prediction to model non-linear cellular dynamics. Furthermore, a transformer-based association network is introduced to learn temporal lineage relationships across multiple frames, enabling robust tracking under complex motion patterns and frequent cell divisions. We evaluate our framework on publicly available datasets, including DeepCell (Moen et al., 2019) and CTC (Maška et al., 2023). The contributions of our work are summarized as follows:

- We introduce CytoTracker, an end-to-end cell tracking framework that integrates YOLOX-based detection, diffusion-based motion prediction, and transformer-based association, effectively handling cell division, and long-term tracking.
- We develop CytoEmbedding, a feature fusion mechanism that incorporates spatial and appearance features by combining image features with cell coordinates, enhancing cell-level representation for improved tracking.
- We conduct extensive experimental evaluations on a diverse set of microscopy and natural image datasets, demonstrating state-of-the-art performance in both tracking and association accuracy. Our method performs favourably compared to the state-of-the-art Trackastra (Gallusser and Weigert, 2024), which requires ground truth masks for each frame, even during inference, while our method does not require such expensive annotations

2. Methodology

Our approach consists of three main components: (1) cell detection, (2) non-linear motion prediction, and (3) cell association as shown in the figure 2.

2.1. Cell Detection

Cell detection is treated as a single-class object detection problem using YOLOX (Ge et al., 2021). For each frame $I_t \in \mathbb{R}^{H \times W \times 3}$, the model outputs bounding boxes $B_t = \{b_1, b_2, \ldots, b_N\}$, where each box is represented as $b_i = (x_i, y_i, w_i, h_i, c_i)$, with (x_i, y_i) being the top-left corner coordinates, w_i, h_i as width and height, and c_i as the confidence score.

2.2. Non-linear Motion Prediction Using Diffusion Model

Given an object trajectory $T = \{B_1, \ldots, B_f, \ldots, B_N\}$, where each bounding box at frame f is represented as $B_f = (x_f, y_f, w_f, h_f)$, the object motion M_f is defined as:

$$M_f = B_f - B_{f-1} = (\Delta x_f, \Delta y_f, \Delta w_f, \Delta h_f)$$

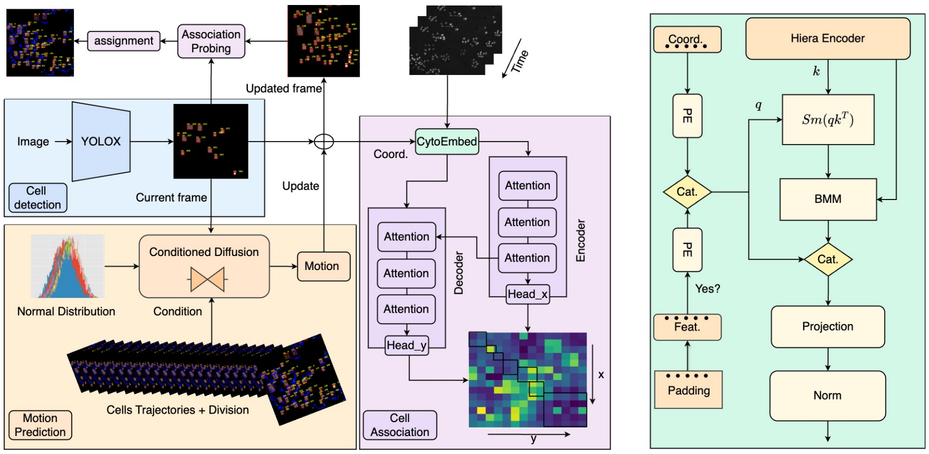

Figure 2: **On Left**: The overall architecture of the proposed CytoTracker, which comprises detection, diffusion-based motion prediction, and association modules. First, we obtain the cell coordinates at the frame level in microscopy videos using a pre-trained YOLOX detector. Next, we introduce a diffusion model to predict cell motion over a short temporal window, which is used to estimate the cell's position in the next frame and guide the cell association. Finally, cell association begins with the CytoEmbedding, which takes estimated cell positions and frame-level image features as input. Then, pairwise associations between cells across frames are performed using an encoder-decoder attention. A candidate graph is then constructed by averaging associations over a sliding window, and the final tracking is refined using either a greedy algorithm or discrete optimization (Kuhn, 1955). **On Right:** The CytoEmbedding extracts image features using a self-supervised pre-trained Hiera (Ryali et al., 2023) model to enhance association learning. This provides image-based, cell-level information through feature fusion.

We model M_f using a diffusion process, which consists of two sub-processes: data-to-zero forward noising process and zero-to-noise backward denoising process.

Forward Process: The data-to-zero process attenuates the clean motion data M_f to zero over time $t \in [0, 1]$, defined as:

$$D_{f,t} = M_{f,0} + t \cdot c$$

where $c = -M_{f,0}$ ensures that $D_{f,1} = 0$.

Backward process: Simultaneously, the zero-to-noise process adds Gaussian noise to the zero data, increasing it to pure noise at $t = 1$:

$$W_{f,t} = \sqrt{t}z$$

where $z \sim \mathcal{N}(0, I)$ is sampled from a normal distribution.

Combined Forward Process: The noisy motion data $M_{f,t}$ at time t is computed as:

$$M_{f,t} = D_{f,t} + W_{f,t} = M_{f,0} + t \cdot c + \sqrt{t}z$$

At $t = 1$, the clean motion data has been fully converted into noise, allowing the reverse diffusion process to predict the future motion of each object.

Bounding Box Prediction The predicted motion M_f is used to update the bounding box in frame f, which is then passed to the cell association module.

$$B_f = B_{f-1} + M_f = (x_{f-1} + \Delta x_f, y_{f-1} + \Delta y_f,\ w_{f-1} + \Delta w_f, h_{f-1} + \Delta h_f)$$

2.3. Association Module

Inspired by Trackastra (Gallusser and Weigert, 2024), we employ a transformer-based association module to improve cell tracking across frames as shown in the Figure 2. For each pair of consecutive frames, we extract an association matrix $\mathbf{A} = [A_{ij}]$, where A_{ij} represents the probability of associating cell i in frame $f-1$ with cell j in frame f. Unlike Trackastra, our model is trained using bounding box crops and hiera (Ryali et al., 2023) image encoder.
Input Acquisition. We acquire input data, consisting of high-resolution time-lapse microscopy videos capturing dynamic cell behavior and cell detections. The raw image sequences provide detailed visual information of the cells over time, while the bounding boxes (coord.) from the detection module define regions of interest in each frame.
CytoEmbedding. The pipeline of the CytoEmbedding is illustrated in Figure 2. This module fuses spatial, temporal, and contextual image features to create unified cell representations for tracking. Input images are first processed by a pre-trained image encoder to extract features. Cell coordinates $\mathbf{p}_i$ are normalized by removing temporal offsets and embedded using a positional encoder, while optional cell features $\mathbf{z}_i$ are concatenated with positional embeddings. An attention mechanism aligns cell features with image features, generating context vectors through weighted summation. These context vectors are fused with the cell features, projected to a lower-dimensional space, and normalized, producing the final unified representation. This enables robust tracking by capturing complex spatial and temporal relationships.

To represent each object effectively, tokens $\mathbf{x}_i$ are constructed by applying Fourier Positional Encodings (Θ) to position features $\mathbf{p}_i$. These encodings are concatenated with shape descriptors $\mathbf{z}_i$ and projected using a linear transformation $\mathbf{x}_i = W_{\text{inp}} \cdot \text{concat}(\Theta(\mathbf{p}_i), \mathbf{z}_i)$. This unifies both spatial and morphological cell-level information.
Encoder-decoder architecture. The encoder processes tokens $\mathbf{x}_i$ through self-attention layers $A^{\ell}_{f}(\mathbf{X}, \mathbf{X}, \mathbf{X})$, incorporating Rotary Positional Embeddings (RoPE) (Su et al., 2024) to capture relative spatial and temporal relationships. This produces contextualized representations $\mathbf{Y}$. The decoder refines tokens by applying cross-attention layers $A^{\ell}_{g}(\mathbf{X}, \mathbf{Y}, \mathbf{Y})$ between the input tokens $\mathbf{X}$ and the encoder output $\mathbf{Y}$. The resulting refined representations $\mathbf{Z}$ are further processed by multi-layer perceptrons (MLPs).

Finally, the association logits $\hat{\mathbf{A}}$, representing the likelihood of associations between objects, are computed as the outer product of the outputs: $\hat{\mathbf{A}} = \text{MLP}(\mathbf{Y}) \otimes \text{MLP}(\mathbf{Z})$, where $\hat{\mathbf{A}}$ captures both spatial and temporal dependencies critical for association tasks. This pipeline ensures efficient and accurate cell tracking by leveraging both learned features and positional relationships.
Association Normalization. We normalize the association logits using a specialized Parental Softmax function (Φ) (Gallusser and Weigert, 2024), ensuring that the sum of association probabilities for each object does not exceed one. This normalization enforces biological constraints, such as a cell not being associated with multiple parent cells simultaneously.

We train the model using a weighted binary cross-entropy loss function. This loss prioritizes associations involving dividing cells and track continuations, which are critical for accurate cell lineage tracking. By weighting these associations more heavily, the model enhances its accuracy and robustness in capturing complex cellular behaviors.

Inference and Linking. To generate global tracking results, we first perform sliding window averaging, where predictions from overlapping temporal windows are averaged to obtain global association scores $\bar{\mathbf{a}}_{ij}$. We then construct a candidate graph where nodes represent objects and edges represent potential associations, discarding edges with low scores to reduce noise. Finally, we apply linking algorithms to derive the final tracks. This involves methods such as Greedy Linking, which iteratively adds the most probable edges while satisfying biological constraints; Linear Assignment Problem (LAP) (Fukai and Kawaguchi, 2023; Jaqaman et al., 2008), which solves an optimization problem to find globally optimal associations; or Integer Linear Programming (ILP) (Gallusser and Weigert, 2024; Malin-Mayor et al., 2023), which enforces stricter constraints to ensure optimal solutions. The final output is a set of cell tracks, representing the trajectories of individual cells over time. These tracks account for cell divisions and migration, providing a comprehensive view of cell behavior in the microscopy videos.

3. Experiments

Experimental Details. CytoTracker is trained in three stages using a single NVIDIA A100 40GB GPU. First, we train YOLOX (Ge et al., 2021) for cell detection. The detected cell positions are then used to train a diffusion model for motion prediction. Finally, we train the association module to link cell instances across frames. For training, we set the window size to 4, the embedding dimension to 768, and the number of encoder and decoder attention layers to 6. The model processes a maximum of 1024 tokens per window, constraining the total number of cells, and is trained with a batch size of 4.

Evaluation Metrics. We evaluate our cell tracking method using False Positives (FP) and False Negatives (FN) to measure incorrect and missed associations. Acyclic Oriented Graph Matching (AOGM) (Matula et al., 2015) quantifies the operations needed to transform the predicted graph into the ground truth. Tracking Accuracy (TRA) normalizes AOGM relative to an empty graph, ensuring a standardized comparison. Furthermore, Association Accuracy (Asso Acc) (Hayashida et al., 2020) assesses the correctness of object associations across frames (Matula et al., 2015). For natural image object detection, we use HOTA, IDF1, AssA, MOTA, and DetA, which measure tracking accuracy, identity preservation, association correctness, and detection performance (Bernardin and Stiefelhagen, 2008; Luiten et al., 2021).

3.1. Results and Analysis

Performance on DeepCell Dataset. Table 1 presents a comparative analysis of state-of-the-art tracking methods on the DeepCell dataset. Our proposed method, CytoTracker, achieves the second-best overall performance, with an AOGM score of 6.4 and a perfect TRA of 1.000. While Trackastra-General (Gallusser and Weigert, 2024) (ILP) achieves the lowest AOGM score (5.8), it requires cell segmentation information during inference. In contrast, CytoTracker operates directly on microscopy images, eliminating the need for segmentation masks during inference, making it more practical for real-world applications. Furthermore, CytoTracker outperforms most baselines in division tracking, achieving a Div F1 score of 0.96, which is on par with Trackastra-General (Greedy) and superior to other competing methods. Additionally, CytoTracker attains an Association Accuracy (AA) of 1.00, matching the highest-performing models. This highlights its robustness in maintaining accurate cell identities across frames, ensuring reliable long-term tracking.

Table 1: Performance comparison of various tracking methods on the DeepCell dataset.

Model	**Linear optim**	**AOGM ↓**	**TRA ↑**	**Division F1 ↑**	**Asso Acc ↑**
Baxter (Magnusson et al., 2015)	Greedy	-	0.997	0.72	1.00
CellTrackerGNN (Ben-Haim and Raviv, 2022)	Greedy	128.3	0.999	0.18	0.93
Caliban (Moen et al., 2019)	LAP	18.1	1.000	0.97	0.99
Trackastra	Greedy	12.2	1.000	0.90	1.00
Trackastra	ILP	7.9	1.000	0.94	1.00
Trackastra-General	Greedy	7.4	1.000	0.96	1.00
Trackastra-General	ILP	5.8	1.000	0.94	1.00
CytoTracker (Ours)	ILP	6.4	1.000	0.96	1.00

Baseline on MOT tasks. Table 2 presents a comparative analysis of various multi-object tracking (MOT) methods on the DanceTrack dataset. Among the existing methods, DiffMOT achieves the best overall performance, attaining the highest HOTA (62.3), IDF1 (63.0), and AssA (47.2), along with strong MOTA (92.8) and DetA (82.5) scores. This highlights the effectiveness of diffusion-based motion prediction in multi-object tracking tasks. Our proposed method, CytoTracker, matches DiffMOT in HOTA (62.3) but outperforms it in IDF1 (63.2) and Ass Acc (48.7), demonstrating superior tracking association accuracy. This improvement is attributed to our transformer-based association module, which effectively captures long-range dependencies between objects, leading to more reliable tracklet associations. Additionally, CytoTracker achieves the highest MOTA (93.0) while maintaining a DetA of 82.5, further reinforcing its robustness in maintaining object identities over time.

Table 2: Performance comparison of various tracking methods on the DanceTrack dataset.

Method	**HOTA ↑**	**IDF1 ↑**	**AssA ↑**	**MOTA ↑**	**DetA ↑**
ByteTrack (Zhang et al., 2022)	47.3	52.5	31.4	89.5	71.6
OC-SORT (Cao et al., 2023)	55.1	54.2	38.0	89.4	80.3
DeepOC-SORT (Maggiolino et al., 2023)	61.3	61.5	45.8	92.3	82.2
DiffMOT (Lv et al., 2024)	62.3	63.0	47.2	92.8	82.5
CytoTracker (Ours)	62.3	63.2	48.7	93.0	82.5

Table 3: AOGM scores for different feature combinations and linkage.

Coord.	**Image**	**Motion**	**Cell Feat.**	**Linkage**	**AOGM ↓**
✓	✗	✗	✗	Greedy	87.9
✗	✓	✗	✗	Greedy	213.4
✓	✓	✗	✗	Greedy	12.1
✓	✗	✗	✓	Greedy	11.9
✓	✓	✓	✗	Greedy	11.9
✓	✓	✓	✗	ILP	6.4
✓	✓	✓	✓	ILP	4.9

The Role of Rich Feature Combinations: Table 3 presents an ablation study exploring the impact of various feature combinations and linkage methods (greedy vs. ILP) on AOGM scores. Using only spatial coordinates (**coord.**) with the greedy linkage results in a high AOGM score of 87.9. Incorporating additional features such as image and motion (**coord. + image feat. + motion**) significantly reduces the AOGM score to 11.9 with the greedy method and to 5.2 when using the ILP method.

Impact of ID Switches and Discovered Objects in MPOT Task: The qualitative results, shown in Figure 3 for CTC challenging dataset Fluo-N2DL-HeLa (Maška et al., 2023), highlight the challenges of non-linear motion tracking and association in the MOT task. A

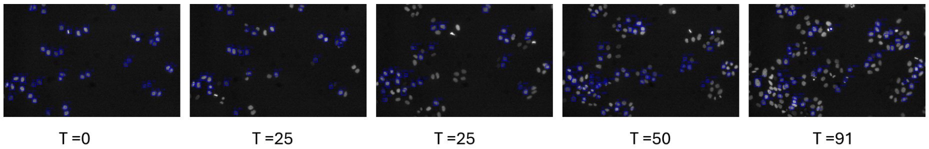

Figure 3: Increasing cell ids due to memory issue.

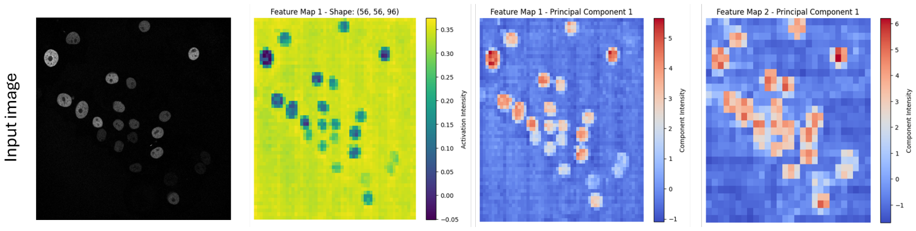

Figure 4: Visualization of image features for fusion with coordinates. The features are extracted using the Hiera encoder and visualized through aggregation, as well as PCA on two-stages (PCA1 and PCA2) for key feature.

key observation is the increasing number of cell IDs, which is attributed to memory limitations during tracking. This results in frequent ID switches, where a single cell is mistakenly assigned multiple IDs over time, disrupting the continuity of tracking. Furthermore, the results indicate an increase in the number of newly discovered objects, particularly in scenarios with complex non-linear motion. While this reflects the model's ability to identify previously undetected cells, it also suggests difficulties in consistently associating objects across frames. These challenges emphasize the critical role of the tracking association module, which ensures accurate object linking across frames, even in complex and occlusion-heavy scenarios.

Image Feature Visualization for Fusion: The visualization of image features from CTC Fluo-N2DH-GOWT1 (Maška et al., 2023) (as shown in the Figure 4 for fusion with coordinates demonstrates the effectiveness of the Hiera encoder in extracting meaningful representations. These features are aggregated and subjected to a two-stage Principal Component Analysis (PCA1 and PCA2), highlighting key components that encapsulate critical information. The resulting features can be interpreted as cell-level features, where the intensity of components corresponds to specific cell information. This approach allows for capturing spatial and morphological details at a granular level, making the extracted features highly relevant for tasks like segmentation and tracking. The qualitative results showcase the ability of these features to encode cell-specific information effectively, reinforcing their utility in downstream tasks.

4. Conclusion

CytoTracker provides a robust foundation, augmented with domain-specific adaptations and significant advancements in association mechanisms and motion modeling. CytoTracker holds the potential to establish a new benchmark for tracking dividing objects in both biological and non-biological domains. Future work will focus on scaling the model to encompass a wider range of datasets and applications, further enhancing its robustness, adaptability, and practical utility for researchers and practitioners alike.

References

Tal Ben-Haim and Tammy Riklin Raviv. Graph neural network for cell tracking in microscopy videos. In *European Conference on Computer Vision*, pages 610–626. Springer, 2022.

Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.

Jinkun Cao, Xinshuo Weng, Rawal Khirodkar, Jiangmiao Pang, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9686–9696, 2022. URL https://api.semanticscholar.org/CorpusID:247763039.

Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9686–9696, 2023.

Y. Chen, M. Rohrbach, Z. Yan, S. Yan, J. Feng, and Y. Kalantidis. Tracking anything with transformers. *arXiv preprint arXiv:2012.03084*, 2020.

Yohsuke T Fukai and Kyogo Kawaguchi. Laptrack: linear assignment particle tracking with tunable metrics. *Bioinformatics*, 39(1):btac799, 2023.

Benjamin Gallusser and Martin Weigert. Trackastra: Transformer-based cell tracking for live-cell microscopy. *arXiv preprint arXiv:2405.15700*, 2024.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021, 2021. URL https://arxiv.org/abs/2107.08430.

Junya Hayashida, Kazuya Nishimura, and Ryoma Bise. Mpm: Joint representation of motion and position map for cell tracking. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3822–3831, 2020. URL https://api.semanticscholar.org/CorpusID:211296729.

Khuloud Jaqaman, Dinah Loerke, Marcel Mettlen, Hirotaka Kuwata, Sergio Grinstein, Sandra L Schmid, and Gaudenz Danuser. Robust single-particle tracking in live-cell time-lapse sequences. *Nature methods*, 5(8):695–702, 2008.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129:548–578, 2021.

Weiyi Lv, Yuhang Huang, Ning Zhang, Ruei-Sung Lin, Mei Han, and Dan Zeng. Diffmot: A real-time diffusion-based multiple object tracker with non-linear prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19321–19330, 2024.

Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3025–3029. IEEE, 2023.

K. E. G. Magnusson, J. Jaldén, P. M. Gilbert, and H. M. Blau. Global linking of cell tracks using the viterbi algorithm. *IEEE Transactions on Medical Imaging*, 34(4):911–929, 2015.

Caroline Malin-Mayor, Peter Hirsch, Leo Guignard, Katie McDole, Yinan Wan, William C Lemon, Dagmar Kainmueller, Philipp J Keller, Stephan Preibisch, and Jan Funke. Automated reconstruction of whole-embryo cell lineages by learning from sparse annotations. *Nature biotechnology*, 41(1):44–49, 2023.

Martin Maška, Vladimír Ulman, Pablo Delgado-Rodriguez, Estibaliz Gómez-de Mariscal, Tereza Nečasová, Fidel A Guerrero Peña, Tsang Ing Ren, Elliot M Meyerowitz, Tim Scherr, Katharina Löffler, et al. The cell tracking challenge: 10 years of objective benchmarking. *Nature Methods*, 20(7):1010–1020, 2023.

Pavel Matula, Martin Maška, Dmitry V Sorokin, Petr Matula, Carlos Ortiz-de Solórzano, and Michal Kozubek. Cell tracking accuracy measurement based on comparison of acyclic oriented graphs. *PloS one*, 10(12):e0144959, 2015.

Erick Moen, Enrico Borba, Geneva Miller, Morgan Schwartz, Dylan Bannon, Nora Koe, Isabella Camplisson, Daniel Kyme, Cole Pavelchek, Tyler Price, et al. Accurate cell tracking and lineage construction in live-cell imaging experiments with deep learning. *Biorxiv*, page 803205, 2019.

Owen M O'Connor, Razan N Alnahhas, Jean-Baptiste Lugagne, and Mary J Dunlop. Delta 2.0: A deep learning pipeline for quantifying single-cell spatial and temporal dynamics. *PLoS computational biology*, 18(1):e1009797, 2022.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International Conference on Machine Learning*, pages 29441–29454. PMLR, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

P. Sun, J. Cao, Y. Jiang, R. Zhang, Y. Xiong, and P. Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

Changcheng Xiao, Qiong Cao, Yujie Zhong, Long Lan, Xiang Zhang, Zhigang Luo, and Dacheng Tao. Motiontrack: Learning motion predictor for multiple object tracking. *Neural Networks*, 179:106539, 2024.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer, 2022.

Yudong Zhang and Ge Yang. A motion transformer for single particle tracking in fluorescence microscopy images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 503–513. Springer, 2023.