A detector-independent quality score for cell segmentation without ground truth in 3D live fluorescence microscopy

Jules Vanaret, Victoria Dupuis, Pierre-François Lenne, Frédéric Jp Richard, Sham Tlili, Philippe Roudot

A detector-independent quality score for cell segmentation without ground truth in 3D live fluorescence microscopy

Jules Vanaret, Victoria Dupuis, Pierre-François Lenne, Frédéric Richard, Sham Tlili, Philippe Roudot

***Abstract*—Deep-learning techniques have enabled a breakthrough in the robustness and execution time of cell segmentation algorithms for fluorescence microscopy datasets. However, the heterogeneity, dimensionality and ever-growing size of 3D+time datasets challenge the evaluation of measurements. Here we propose an estimator of cell segmentation accuracy that is detector-independent and does not need any ground-truth nor priors on object appearance. To assign a segmentation quality score, our method learns the dynamic parameters of each cell to detect inconsistencies in local displacements induced by segmentation errors. Using simulations that approximate the dynamics of cellular aggregates, we demonstrate the score ability to rank the performance of detectors up to 40% of false positives. We evaluated our method on two experimental datasets presenting contrasting scenarios in density and dynamics (stem cells nuclei in organoids and carcinoma cells in a collagen matrix) using two state-of-the-art deep-learning-based segmentation tools (Stardist3D and Cellpose). Our score is able to appropriately rank their performances as reflected by accuracy (centroid matching) and precision (segmentation overlap).**

***Index Terms*—Image segmentation, Error analysis, Biological cells, Microscopy, Fluorescence, Stochastic processes, Particle tracking, Dynamics, Image motion analysis, Biophysics.**

I. Introduction

RECENT progress in fluorescence microscopy has enabled high-resolution volumetric imaging of complete cellular systems in their physiological context with minimal phototoxicity, fast sampling, and near-isotropic resolution [1]–[3]. Cell detection methods based on deep learning have been a breakthrough toward the automated quantification of such terascale datasets where the shape and textures of objects of interest can vary widely depending on the cell type, the microscope and the environment [4]–[6]. However, their performances are still limited by the difficult annotation of cellular heterogeneity in complex three-dimensional datasets as well as inhomogeneous signal across the volume. Furthermore, their performances can be challenging to predict, as even recent and widely used approaches output different results on the same sample (Figure 1a,b). While an unbiased evaluation and comparison of those different tools is already a difficult task in two dimensions, the task is virtually impossible in 3D+time sequences as data size increases quadratically compared to a 2D segmentation.

In this paper, we propose a novel score estimator to evaluate the accuracy of cell segmentation tools in live cell fluorescence microscopy imaging. In a nutshell, our score exploits the variation in temporal consistency to predict detection errors while remaining fully independent from the type of detector used (Figure 1c). The idea behind the approach comes from the observation that segmentation errors provide a time-varying response that is different from the usual dynamics. For example a cell cluster can be mis-detected as a single cell, a cell mis-detected as two cells, and a false positive detected in the background. To measure those inconsistencies, we first infer the parameters of cellular dynamics for each trajectory hypothesis, then, we evaluate for each detection the stability of the optimal set of trajectory by combining discrete optimization and a statistical resampling of the trajectory-to-detection likelihood. This study builds upon our previous work on trackability inference in the context of diffraction limited particles [7] with key differences: the detection is here a much more challenging task with a broader variety of error types and our approach focuses here on detection quality rather than trajectory.

The paper is organized as follows. We first provide a brief review of quality scores that have been proposed to evaluate detection algorithms [8]–[10]. To our knowledge, no approaches have been proposed for a comparison of detection results without ground truth or *a priori* knowledge on the object structure. Second, we present the design of our score, our stochastic motion modeling for cellular dynamics as well as the combinatorial optimization framework for multiple hypothesis tracking. Third, we study the behavior of our estimator on simulated data, specifically its robustness toward false positives, false negatives as well as error-induced split and merged cells. Fourth, we then demonstrate the performances of our score in predicting local segmentation errors, and overall F1 score on experimental datasets: on two 3D datasets (a challenging two-photon 3D live imaging of nuclei in organoid [11] specimens and a reference data of human breast carcinoma cells from the Cell Tracking Challenge [12]), we show that our score is able to rank the performances of both pre-trained and trained cell detectors accurately. Our data also suggest that the heterogeneity of the underlying cellular dynamics may impact ranking accuracy. We finally discuss how the idea of our approach can be further generalized.

The project leading to this publication has received funding from the "Investissements d'Avenir" French Government program managed by the French National Research Agency (ANR-16-CONV-0001) and from Excellence Initiative of Aix-Marseille University - A*MIDEX

J. Vanaret, F. Richard, and P. Roudot are with Aix-Marseille Université, CNRS, I2M UMR 7373, Turing Centre for Living systems, Marseille, France.

J. Vanaret, V. Dupuis, P-F. Lenne, and S. Tlili are with Aix-Marseille Université, CNRS, IBDM UMR 7288, Turing Centre for Living systems, Marseille, France.

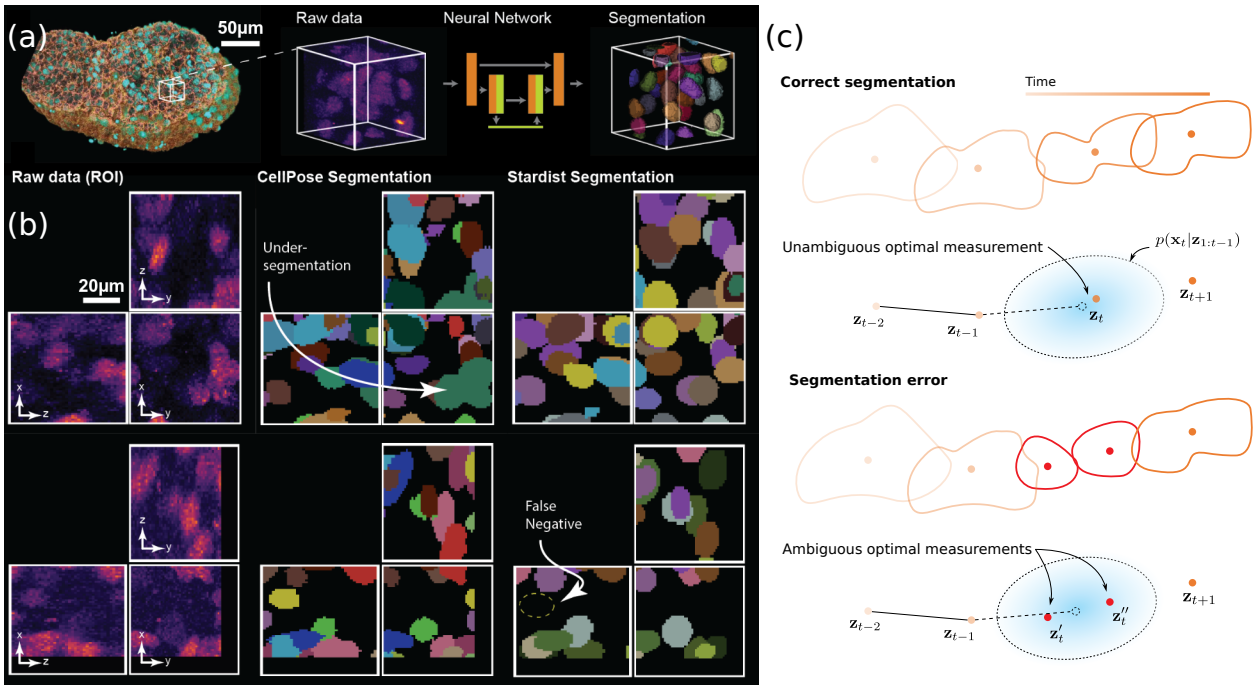

Fig. 1. The density and heterogeneity of cellular objects challenge the comparison and validation of leading-edge segmentation approaches based on deep neural networks. (a) 3D volumetric rendering of a gastruloid imaged with two-photon microscopy along with a maximum intensity projection (MIP) rendering of a region of interest (ROI) and its associated segmentation. (b) Two ROIs rendered as orthogonal MIPs showing different artifacts ranging from the detection of nuclei clusters as a single nucleus (under-segmentation) to false negatives. (c) Our accuracy score uses temporal consistency to evaluate the likelihood of errors. Top: Without segmentation errors, the detector provides a single measurement compatible with the cell dynamics. Bottom: A segmentation error, here a split, will result in two measurements with ambiguous associations to the trajectory.

II. Related Work

A few wide-ranging studies [9], [12] have used manually-annotated data to analyze the performances of multiple cell segmentation algorithms. The metrics used for comparison typically distinguishes between the ability to segment a cell, using a metric of the overlap between estimated and annotated cell mask, and the ability to detect a cell, using a measure of object instance matching accuracy. On the one hand, a study focusing on cell morphogenesis will need a precise matching of the measured contour with the ground truth. On the other hand, a study analyzing cell count, lineage and cycle will require high detection accuracy against missed detection, false positives, split and merge artifacts. Since the seminal challenge on cell detection and tracking algorithms [12], deep learning approaches have enabled a breakthrough in both cell segmentation and detection. In their detailed and unbiased comparison of the capacity of conventional and deep-learning-based approaches to measure biologically relevant metric, Caicedo and colleagues [9] have shown that deep learning approaches are indeed performing better overall. However, their results also show that benchmarking techniques remain more important than ever. For example, while the count of false negatives obtained with U-net [4] is almost twice lower than with conventional adaptive thresholding approaches, the amount of cells artificially split by the neural network is higher than a threshold-based approach and a random-forest-based classifier. They also show that the performances of deep neural networks can vary significantly from one network structure to another or depending on the amount of annotations available. Importantly, while the relative performance of a conventional approach with respect to a specific task can be intuitively interpreted from the parameters and features that are used, the best purpose of a neural network can only be done with comprehensive benchmarking of its breakpoint. As such, comparing detectors remains a critical task in the routine use of cell segmenter.

Whether it is for benchmarking or training, annotation in 2D images is a time consuming process that becomes even longer, challenging and bias-prone in datasets of larger dimension. First, each Z plane and time point must be annotated to match the degree of accuracy in the entire volume. For example, in [9], authors measure 50 hours to annotate 11500 objects in 2D static images and efficiently train the network. Reproducing this process for a 3D dataset (with typically 5-10 2D Z-planes per object) would take up to 500 hours, even when ignoring the need to annotate all time points. Second, the 3D nature is difficult to interpret for the user. Correspondences must be made between Z-plane and time point, a non-trivial task when both images are not seen at the same time. 3D rendering is also prone to bias, as two different views may not result in the same visual interpretations [13]. To overcome those challenges, several papers have been focusing on improving annotation techniques for multidimensional datasets. In [14], authors propose to solve an inverse problem to recreate a 3D annotated volume from a collection of 2D annotations on maximum intensity projections. However, this approach is designed for sparse objects (e.g. vessels images) and poorly suited for the segmentation of the cluttered scene of cellular

aggregates. In [15], authors experiment with the eye-tracking capacity of a virtual reality headset to annotate cell location and displacement from mere observation. While the tracking is seamless and robust to wavering attention by the annotator, this remains impractical for a large number of cells and ill-adapted to very dense scenarios. As such, there is to our knowledge no efficient solution to the annotation of 3D+time dataset, a challenge when comparing various approaches.

Considering the ever-growing amount of data produced by next generation microscopy and its multidimensional nature, a few approaches have been developed for annotation-free quality control through the use of prior information on the object shape. First, goodness-of-fit analysis is routinely used to evaluate detection quality when the objects follow a simple template. For example, in a study on high-throughput super-resolution imaging, Beghin and colleagues [8] use chi-square maps to control point-source detection across all the imaged wells. Since the cell intensity does not follow such a strict diffraction-limited model, those approaches cannot be used in our scenario. Fehri et al [16] proposed to alleviate this limitation with a graph-based criteria that characterize contour detection quality through a criteria on object intensity smoothness and object-to-background transition. More recently, Audelan et al [10] proposed a probabilistic approach score for the validation of generic segmentation tools based on the same intensity-base assumptions. However, those quality scores are only designed to evaluate the precision of the segmented contour as opposed to the accuracy of cell detection, a fundamental aspect of quantitative cell biology. More importantly, all those approaches make assumptions on object appearances that may not hold in experimental acquisition and especially in the challenging case of interest: low signal-to-noise ratio can corrupt intensity and blur the line between background and object, high scattering can modify this intensity across the object itself and low optical section typically makes the evaluation of contour quality more difficult in the axial orientation. To tackle those limitations, our contributions in the paper focus on a detector-independent quality score that makes no assumption on the object appearance, but rather on its dynamics. In particular, we focus on the capacity of this score to discriminate between false positive, negative, merging and splitting events. To our knowledge, this approach and the study of its performance is original and will hopefully pave the way to further investigation in this field.

III. Methods

Our approach aims at detecting segmentation errors through the dynamic footprint of objects' dynamics. In this section, we first present the framework of Bayesian filtering to learn motion parameters, then we explain how we exploit ambiguities in local trajectories-to-measurement associations to detect potential errors.

A. Inferring motion parameters

We use the Bayesian filtering framework to learn the parameters of each object's dynamics in a temporally greedy fashion. In this formalism, the state of an object at frame $t \in \mathbb{N}$ is represented by $x_t \in \mathbb{R}^N$, with N the number of state variables in our model. For example,

$$x_t = (x(t), y(t), z(t), \dot{x}(t), \dot{y}(t), \dot{z}(t)) \tag{1}$$

can be used to model the state of a point-like particle in 3D moving in a directed fashion. Under the Markovian chain hypothesis that x_{t+1} can be determined exclusively from x_t, one can propose a dynamical model $f : \mathbb{R}^N \times \mathbb{N} \to \mathbb{R}^N$ for x_t known up to some precision represented by $w_t \in \mathbb{R}^N$ defined as a realization of a random variable that represent the model (or process) noise:

$$x_{t+1} = f(x_t, t) + w_t. \tag{2}$$

w_t could represent deviations from the reality of an over-simplistic model, or real stochastic terms in the equation of the dynamics, like Langevin forces [17]. The complete state variable x_t is hidden and information about the object can only be obtained via the measurement of variable $z_t \in \mathbb{R}^M$, with $M \leq N$ (e.g position but not velocity in the example above). The process of measurement is modeled by a function $h : \mathbb{R}^N \times \mathbb{N} \to \mathbb{R}^N$ corrupted by $v_t \in \mathbb{R}^M$ defined as a realization of a random variable modeling a measurement noise such that:

$$z_t = h(x_t, t) + v_t. \tag{3}$$

In biological applications, v_t could represent measurement errors due to optical limitations (e.g particles below the diffraction limit appearing as point-spread functions) or systematic errors in detection algorithms.

Considering that the prior probability density function (PDF) $p(x_{t+1}|z_{1:t})$, where $z_{1:t}$ denotes all measurements of a single object up to time t, is known, the Bayesian filtering equation provides the posterior PDF $p(x_{t+1}|z_{1:t+1})$, with:

$$p(x_{t+1}|z_{1:t+1}) \propto p(z_{t+1}|x_{t+1}) \int p(x_{t+1}|x_t) p(x_t|z_{1:t}) dx_t \tag{4}$$

In a multiple unlabeled target tracking (MTT) framework, a common approximation consists in iteratively associating each measurement z_t to its most likely state x_t.

Simplifying assumptions are often made to make the posterior PDF tractable [18]. Notably the models f and h for the dynamics and measurement are assumed to be linear, i.e $f(x_t, t) = Fx_t$ and $h(x_t, t) = Hx_t$, and the noises variables w_t and v_t are assumed to be sampled from Gaussian laws. In this case, $p(x_{t+1}|z_{1:t})$ and $p(x_{t+1}|z_{1:t+1})$ are normally distributed too. Kalman filters give an optimal [19] estimation of the parameters of the dynamics by providing a recurrent formula to compute the prior and the posterior PDFs at each frame [20]. Kalman filters can be implemented efficiently for thousands of objects in parallel, as is routinely the case in biological imaging, e.g endocytic events, microtubule polymerization at the molecular level and developing embryos at the cellular level.

B. Detecting motion inconsistencies

One of the key challenges we face in inferring ambiguities is that a measurement with a lower trajectory-to-measurement

likelihood does not necessarily mean that the segmentation is false (if the cell changes shape for example). In order to assess the presence of false positive, merging or splitting errors (leaving isolated false negative aside for now), we must estimate the likelihood of association between a measurements and neighboring trajectories and test if the optimal solution is unique, i.e. if other local combinations have a significantly lower likelihood. In theory, this would require the filtering of every possible sequence of measurements, which would grow exponentially even with efficient pruning of hypotheses. As such, similarly to MTT approaches, the comparison is carried out on a per-frame basis. We iteratively solve a one-to-one bipartite graph assignment problem between current track segment ends at frame t and detections at frame $t+1$ through the resolution of a linear assignment problem:

$$\begin{aligned} &\underset{a_{ij}}{\operatorname{argmin}} \sum_{i\in\Omega, j\in D_{t+1}} c_{ij}a_{ij}, \\ &\text{s.t} \sum_{i\in\Omega} a_{ij} = 1, \quad \sum_{j\in D_{t+1}} a_{ij} = 1 \end{aligned} \tag{5}$$

with Ω the current set of track segments ends, D_{t+1} the set of detections in frame $t+1$, $a_{ij} \in \{0,1\}$ denotes the assignment of track segment end i to detection j (1 if the link is made, 0 otherwise), and the cost of association c_{ij} is the trajectory-to-measurement negative log likelihood $c_{ij} = -\log p(z^i_{t+1}|x^j_{t+1})$. The constraints enforce the one-to-one linking condition in the bipartite graph. In a conventional tracking framework, track segment creation and termination are considered by adding virtual nodes in the bipartite graph, and a gating parameter forces the termination of trajectories that have only a low likelihood of association between trajectory and detection.

Two modifications must be made to allow for the inference of detection errors. First, we need to detect if other spurious or missing detections create ambiguities in the graph of association. The problem is convex and the global optimum can be reached exactly using linear programming without the possibility for error inference. In order to detect ambiguities in track-to-detections associations, we resample the predicted state from the PDF $p(x^i_{t+1}|z^i_{1:t})$ and the new linking assignments gives us a direct way to evaluate locally the stability of the optimal solution. Another modification lies in the gating parameters used to explore new associations. While tracking approaches use the motion prediction to accept or reject measurements into the optimization problem, we need to explore the measurement candidates in an area corresponding to possible measurement errors, including merging and splitting which may be larger than the area defined by motion only. Thus, we use a search radius SR that is both bigger than the prediction error d_{pred}, provided by the covariance of the innovation [20] and segmentation error d_{err}. Similarly, SR should be smaller than the average distance between particles d_0 to ensure that our score does not measure ambiguities from linking a particle to a track it does not belong to. As a result we chose to set SR to an intermediate scale given by the geometric mean of the upper bound d_0 and of whichever scale is largest between d_{pred} and d_{err}

$$SR \sim \sqrt{\max\left(d_{err}, d_{pred}\right) \cdot d_0} \tag{6}$$

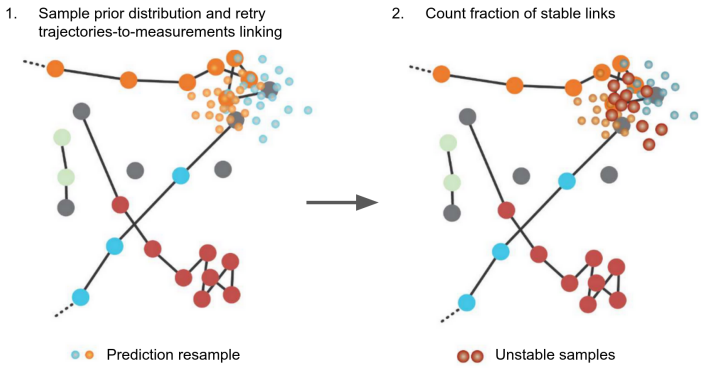

Fig. 2. Our evaluation score uses Monte Carlo resampling of the prior distribution associated with each current track end to measure the instability of the optimal linking previously chosen.

The principle behind our evaluation score is shown on Figure 2. We denote T^{ij}_t our stability score for the given optimal link a^*_{ij} made between frames t and $t+1$, given by

$$T^{ij}_t = \frac{1}{N_s}\sum_{n=1}^{N_s}\left[a^*_{ij} = a^n_{ij}\right], \tag{7}$$

where we perform N_s Monte Carlo resamplings, a^n_{ij} is the newly computed assignment during the n-th resampling, and $[\cdot]$ are Iverson brackets, which equal 1 when the proposition evaluated is true. As such, our local stability score is defined as the ratio between the number of times a link has remained unchanged after resampling and the total number of resamplings. A lower value of the score T^{ij}_t reflects a larger instability in the optimal assignment. As such, a merged, split and false positive detection will create an additional detection candidate that will create an ambiguity captured by the score. While the present work does not model the case of an isolated false negative, our results show that this simple approach can already capture a large array of detection errors. In the experiments, we set $N_s = 100$.

C. Implementation

We based our implementation on the well-established open-source tracking software u-track [7], [18] to take advantage of its efficient Bayesian-filtering framework. This software has been used to track morphologically and dynamically diverse cellular structures. Its flexibility allows the modeling of multiple types of states and associated cellular dynamics in parallel, e.g piecewise Brownian and diffusive dynamics as can be observed with microtubules. Initialization of the Kalman filter parameters is tackled via a forward-backward tracking scheme.

IV. Evaluation metrics

When comparing detections, having a ground-truth allows one to precisely quantify which true objects have been detected

(TP), and which ones have been missed (FN) or correspond to false detections (FP). The F1 score, also called the Dice score, is a popular choice for a universal score taking all error types into consideration [9]. It is defined as the harmonic mean of precision and recall, and reads

$$F1 = \frac{2TP}{2TP + FN + FP}. \tag{8}$$

Note that it is monotonously related to the accuracy metric. If an object can be represented by a single point, e.g if it is point-like by nature or if its centroid gathers most of its positional information, these quantities are computed in a one-to-one assignment framework, in which a detection is consider a TP if it is sufficiently close (up to a threshold distance τ_c) to a point in the ground-truth. We will call this metric the centroid-based F1 score $\mathrm{F1}_c\,(\tau_c)$. Points that are not matched in the detected and ground-truth datasets are considered FPs and TPs respectively. When objects are detected as volumes as opposed to points, they can be made of several pixels/voxels or be represented by enclosing volumes (e.g bounding boxes). The correspondence between two individual objects can be assessed by computing the Intersection over Union score from their volume

$$IoU = \frac{V_D \cap V_{GT}}{V_D \cup V_{GT}} \tag{9}$$

with V_D and V_E the volumes defined by a detected and a ground-truth object respectively. The F1 score can then be computed, in which case a detected object is considered matched to a ground-truth object (and thus a TP) if their IoU is above a certain threshold τ_{IoU} (usually above 0.5). This is again determined in a one-to-one assignment framework. We will call this metric the IoU-based F1 score $\mathrm{F1}_{IoU}\,(\tau_{IoU})$.

V. Validation on synthetic data

Here we want to measure the performances of the detection quality score in a variety of scenarios to establish its capacities and breaking points. Since our method relies strongly on dynamic information and is dedicated to cell or nuclei segmentation, we use simulations that reproduce the motions present in cellular aggregates. We then use these simulations to test the robustness of our quality score with respect to targeted types of errors that reflect segmentation errors made by state-of-the art detection algorithms.

A. Simulation

1) Model: We implemented a multi-particle simulator mimicking the two main dynamical features observed in real cellular movement, namely cells non-interpenetrability and active behavior leading to heterogeneous dynamics. The framework used is inspired by well-established and interpretable active-particle models [17], [21]–[23], where each cell is represented by a point-like particle subject to forces of different microscopic origins. This framework has been successfully used to recreate many complex biological phenomena, like cell-sorting, convergent extension, or active cell jamming [24]. Despite similarity, the notations that follow for acceleration and forces are not related to the stochastic filtering notations introduced in the previous section. We nevertheless elected to keep the standard notations for the sake of readability. The equation of motion, expressing the acceleration $\ddot{x}_i$ for a particle i of mass m_i at position x_i, reads

$$m_i\ddot{x}_i = F_i^{drag} + \sum_{j \text{ neigh.}} F_{ij}^{att-rep} + F_i^{rand}, \tag{10}$$

where the terms on the right-hand side are the forces exerted on the particle, namely viscous drag, attraction-repulsion interactions with neighboring particles, and stochastic forces inducing active behaviors. The combination of all forces is represented on Figure 3a.

First, the particles are subject to a viscous drag force due to the dissipation in cell-cell junctions and in adhesion remodeling [21], with

$$F_i^{drag} = -\gamma_i \left(\dot{x}_i - \langle \dot{x} \rangle_{\aleph_i}\right), \tag{11}$$

where γ_i is an effective friction between neighboring cells, and $\langle \dot{x} \rangle_{\aleph_i}$ is the ensemble averaged velocity of the nearest neighbors $\aleph_i$ of the cell i. Second, to model cell-cell adhesion while preventing cell inter-penetrability, we add Lennard-Jones-type (sticky spheres) spherical forces representing short-range repulsion and mid-range attraction. In our implementation, we approximate the typical Lennard-Jones force profile with a piecewise linear profile inspired by [23], such that the force exerted by a particle j at distance r of i reads

$$\begin{aligned} F_{ij}^{att-rep} &= \varepsilon\left(f_{att}(r; r_i, r_j) - \alpha f_{rep}(r; r_i, r_j)\right) \\ f_{att}(r; r_i, r_j) &= \max(0, r - (r_i + r_j)) \\ f_{rep}(r; r_i, r_j) &= \max(0, r_i + r_j - r) \end{aligned} \tag{12}$$

where r_i and r_j are the radii of particles i and j, ε is the strength of the interaction, and α is a dimensionless parameter that can be used to tune the relative importance of attraction and repulsion. Using a piecewise linear profile improves stability and leads to easier initialization of the simulation while not changing the overall dynamics. The force is set to 0 above a distance $r_{max} = \beta(r_i + r_j)$, with β a proportionality constant ensuring that cells do not interact with their next-nearest neighbors or with neighbors that are too far.

Finally, to drive heterogeneous dynamics and to prevent particles from rapidly reaching a jammed configuration [22], we add a stochastic force term following an isotropic Ornstein-Uhlenbeck process without drift [24]–[26]. Let $F_{i,k}^{rand}$ be the k-th component of the force (e.g $k \in \{x, y, z\}$ in 3D). It satisfies

$$dF_{i,k}^{rand}(t) = -F_{i,k}^{rand}(t)/\tau_i^p dt + f_0 dW_i(t), \tag{13}$$

with persistence time τ_i^p and $f_0 dW_i(t)$ a Wiener process of variance $f_0^2 = 2dD_i$, with d the number of spatial dimensions and D_i the effective diffusion coefficient of particle i. This leads to the magnitude of the stochastic force having a variance of $\sigma^2 = 2dD_i\tau_i^p$. This can be interpreted as a classical Langevin force [17] whose direction can vary randomly in all directions, but with added temporal persistence, resulting in correlation of the magnitude and of the direction of the force over timescales of order τ_i^p. Biologically, these Langevin-type forces can be related to the motility forces associated

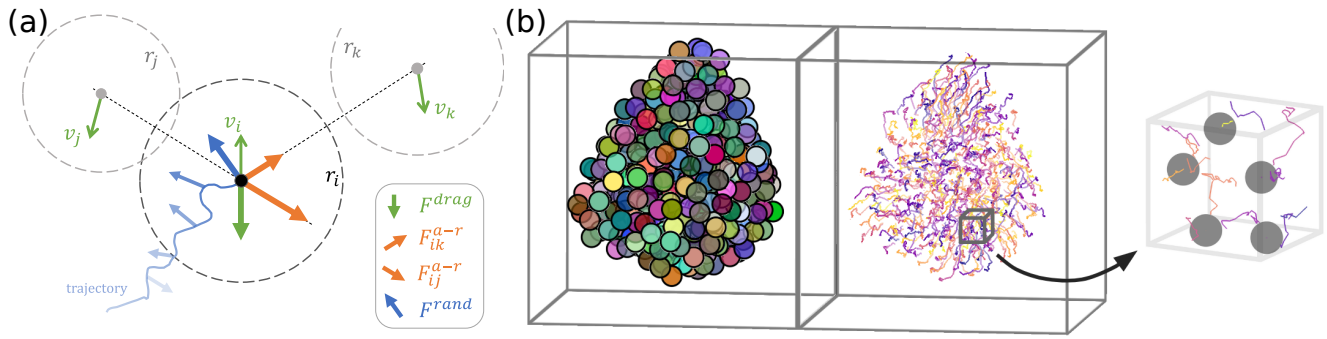

Fig. 3. We use simulations to measure the performances of our detection quality score in a variety of scenarios to establish its capacities and breaking points. (a) In our simulation, particles are subject to three forces: drag due to dissipation at cell-cell interfaces, attraction/repulsion due to non-interpenetrability of the membranes and nuclei, and stochastic Langevin-like forces creating active motion and heterogeneous dynamics. (b) Simulated particles form dense aggregates and show complex dynamics (left: sphere rendering, right: trajectories).

with cell protrusions [24]. Notably, it is easy to modulate the dynamical behavior of cells by varying the individual values of τ_i^p: particles with low persistence times (compared to the observation timescale) move in a diffusive fashion reminiscent of Brownian motion (the Langevin force corresponds to the limit $\tau_i^p \to 0$), whereas cells with high persistence times can move in a quasi-directed manner during timescales of order τ_i^p.

2) Implementation: In our data, the dynamics is over-damped (friction dominates) such that forces exerted on a cell effectively balance at all times. This motivates us to neglect the inertia term in the equation of motion, which becomes a first-order ordinary differential equation in the position of the particles [21], [23]

$$\dot{x}_i = \langle \dot{x} \rangle_{\aleph_i} + \frac{1}{\gamma_i} \left(\sum_{j \text{ neigh.}} F_{ij}^{att-rep} + F_i^{rand} \right). \qquad (14)$$

The term $\langle \dot{x} \rangle_{\aleph_i}$ is computed at t using the velocities at the previous time point $t - \Delta t$ to keep the system of equations separable. This would require us to choose a very low value for Δt, so we use Heun's integration scheme (of order 2) to compensate for this and keep a reasonable value of the time step [23]. To compute the attraction-repulsion forces, we build a k-d tree with the positions of the particles at each time point, which is a data structure that allows for fast lookup of nearest-neighbor relationships and computation of pair-distances. Because these forces vanish beyond a certain distance, the pair-distance matrix is highly-sparse and thus can be computed efficiently. The simulator is implemented in *Python3* and multiprocessing is used with *Numba* to take advantage of the separability of the system of equations at each time point.

The simulation produces sphere-like aggregates of densely packed particles, as shown on Figure 3b. We add a small amount of heterogeneities in the radii of the particles by sampling them from a normal distribution of mean $d_0/2$ and standard deviation $d_0/20$, truncated at $d_0/2 \pm d_0/20$. This leads to particles having an average radius $r_i = d_0/2$ and separated from their immediate neighbors by a distance d_0.

3) Simulation of point-like segmentation errors: We produce typical segmentation errors like false negatives (FN), false positives (FP), splits, and merges by corrupting the positions of the simulated particles at each frame, as illustrated on Figure 4. To add random FN errors at a given rate ρ_{FN} from a simulation with N particles, we draw $\rho_{FN} N$ particles in a uniform manner (every random choice in this section is done following a uniform distribution) at each frame and remove them. For FP errors at a given rate ρ_{FP}, we choose $\rho_{FP} N$ positions uniformly inside the volume of the particles at which we add fake particles. For merges at a rate ρ_{me}, we compute at each frame all pairs of particles that are immediate neighbors from one another, we take $\rho_{me} N$ pairs, remove the associated particles and add fake particles at their barycenter. For splits at a rate ρ_{sp}, we choose $\rho_{sp} N$ particles, and for each one, with radius r_i at position x_i, we choose a random 3D vector u of magnitude $r_i/2$. We then remove the particle and add two fake particles at positions $x_i \pm u$.

B. Behavior of our score under controlled conditions

We simulated $N = 400$ particles for a duration $T = 50\Delta t$, with Δt the integration time step, chosen so that the average displacement between frames matches what we observe in our data (illustrated on Figure 5). Particle positions were initialized by choosing N random points uniformly in a sphere whose radius is such that the density inside the sphere is below that of random sphere packing. This leads to particles expanding outwards before the aggregate settles to its equilibrium radius, after which particles are on average at distance d_0 from their nearest neighbors (Figure 5a). The expansion phase of the dynamics was removed in our experiments. Particles are simulated with infinite lifetimes. To compute our score, we set the search radius SR using our heuristic on the length scales of the different phenomena. Here, the average displacement $d_{dyn} \sim d_0/10$, with d_0 the average nearest-neighbor distance. By design, merge and splits errors produce an average displacements of $d_0/2$ and $d_0/4$ respectively. We thus set

$$SR = \sqrt{(d_0/2) \cdot d_0} \sim 0.7 d_0. \qquad (15)$$

We first study the behavior of the score in conditions corresponding to adding a single error type and varying its rate. While this is an unrealistic scenario that may not reflect the behavior of the score in experimental conditions, this approach

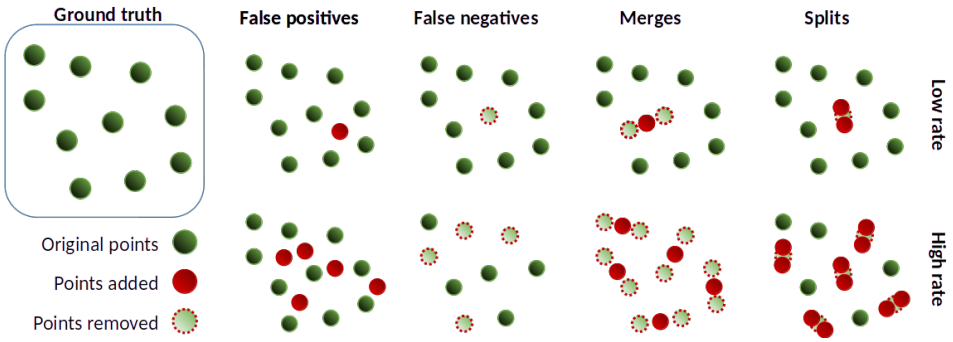

Fig. 4. Illustration of the different types of errors introduced in the simulations. Each error type is shown in a low and a high error rate scenario.

allows us to unbiasedly test the strength and weaknesses of our method. The results are shown on Figure 6a, where we plot the evolution of our score along the centroid-based F1 score $\mathrm{F1}_c\,(\tau_c)$ against error rates. For merge and split conditions, the ground-truth $\mathrm{F1}_c\,(\tau_c)$ is plotted for the complete range of possible values for τ_c, showing that the true performance heavily varies depending on the choice of the matching threshold. The darker $\mathrm{F1}_c$ curve in the middle is obtained for a value τ_c close to the typical scale of the respective events. The dimmer curves correspond to the extremal curves of $\mathrm{F1}_c\,(\tau_c)$ when $\tau_c \to 0$ and $\tau_c \to \infty$. FN and FP events do not have a typical scale associated with them since they will always be detected in those uniform scenarios. We show that for merges, splits and FP errors, our score follows the trend of the true F1 score, both in its monotonicity and in qualitative proximity to its value. For FP errors, we notice that our score is weakly responsive for rates below 15%. This is coherent with our previous works [27] where tracking performances under a varying rate of FP exhibit a similar regime transition with an inflexion point that decreases as object velocity increase. This suggests that FP detection performs better when the density of errors starts to significantly impact tracking. For FN errors, the score shows little sensitivity. This is unsurprising since an isolated false negative will not be creating dynamic instability. As such, while we do not have an exact match with a given accuracy score, our score is capable of ranking those simulated detection approaches without ground-truth in three scenarios.

To study conditions closer to realistic cases, where multiple type of error arise jointly, we plot on Figure 6b the evolutions of our score and of $\mathrm{F1}_c$ for different rates of both FN and FP errors. When corrupting point positions, we first added FN errors before FP to ensure that no FP error added would be canceled by a FN error. We use $\tau_c = SR$ to compute $\mathrm{F1}_c$. We observe that the F1 score decreases significantly with the increase of the FN and FP rates, which is consistent with an increase in the number of errors. Again, we observe that our score is less sensitive to FN than FP errors, as it significantly decreases when increasing the FP rate at a given FN rate. On the other hand, Figure 6c shows that fixing the FP rate and varying the FN rate yields small variations of the score. For small or intermediate values of the FP rate, Spearman's coefficient (defined as the Pearson correlation coefficient applied to the set of rank of each values, instead

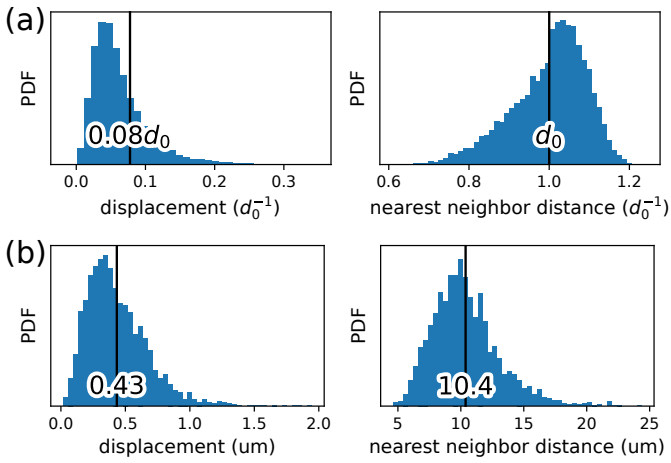

Fig. 5. Distribution of frame-to-frame displacement and nearest neighbor distance in our simulation (a) and in real experimental data (b). Vertical black lines are centered and annotated with the average value of the distributions.

of the values themselves) between the score and true F1 indicates that the ranking property is conserved, but at high FP rates (more than 40%), the score increases when the F1 score decreases (negative Spearman's coefficient). Our score reaches its descriptive limit in this extreme scenario because no trajectories can be formed, even in-between a few time points. As such, no ambiguities in trajectory-to-detections can be measured. Considering the large percentage of FP and FN required to reach its breakpoint, we are confident that our score can properly rank detectors, especially since such a large rate of false negatives or false positives can be detected by mere visual inspection.

VI. Application to Experimental Data

We demonstrate the usefulness of our approach on two live volumetric imaging datasets. The first one shows fluorescent nuclei of mouse embryonic stem cells forming an organoid which we imaged with two-photon microscopy, and is our primary dataset for the application of our works. It exhibits high cellular density and heterogeneities in both cells dynamics and appearance. The second one is a reference dataset from the Cell Tracking Challenge [12] and shows human breast carcinoma cells embedded in collagen. It exhibits sparser cellular

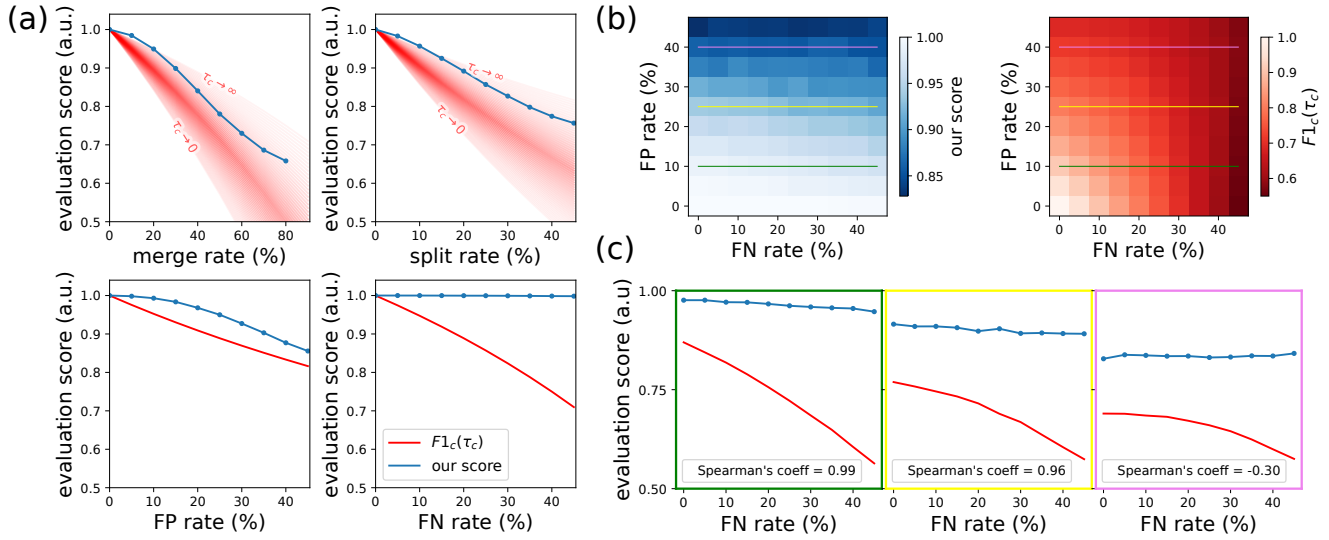

Fig. 6. Our evaluation score responds strongly to simulated merging, splitting and false positive errors and is sensitive to false negatives in scenarios with mixed error types. (a) When a single error type is introduced in simulated data, our score is well correlated with the F1 score except for false negatives, for which it is unresponsive by design. For merge and split conditions, the threshold distance is varied to represent the possible range of F1 score profiles. (b) In mixed FN/FP conditions, our score shows a higher sensitivity to false positives than false negatives, and (c) for high density of false positives, its ranking property can fail. Plots in (c) display the range of values covered by the colored lines in (b).

density but higher heterogeneities in both cells dynamics and appearance, with notably elongated cells.

A. Organoid cells dataset

To decrease the effective visible cell-density, we made the organoids from a mixture of two mouse embryonic cell lines, with initially 50% of non-fluorescent cells (cell line ES-E14TG2a) and 50% of cells endowed with a fluorescent reporter of the gene Brachyury (cell line T-Bra-GFP/NE-mKate2), which makes the cell nucleus fluorescent when the gene is expressed. Biphoton imaging of organoids was performed in a chamber maintained at 37°C, 5% CO2 with a humidifier using a Zeiss 510 NLO (Inverse - LSM) with a femtosecond laser (Laser Mai Tai DeepSee HP, 900 nm) with a 40 x/1.2 C Apochromat objective. A z-stack acquisition of a 50 microns thickness is performed every 5 minutes and every 1 micron. Lateral pixel size is 0.62 microns. A volume from the movie is shown on Figure 7a.

Stardist3D [5] and Cellpose [6] are two open-source state of the art segmentation tools based on neural networks. Stardist3D learns for each voxel a “blueprint” of the star-convex representation of the object it belongs to. Cellpose learns the vector field representing each object as a basin of attraction which can be followed to the object center. Stardist3D is usually more appropriate for ellipsoidal objects in dense environments, and Cellpose often performs better for complex shapes and intensity heterogeneities [28]. Using a custom ground-truth segmentation of this dataset, we first compare Stardist3D and Cellpose using the F1 scores on centroid and on IoU as defined in the Metrics Section. For the sake of emulating a real-life use of segmentation tools in the context of biology, we tested off-the-shelf pre-trained neural networks models for each tool [29], [30] (several pre-trained models are made available by the Cellpose maintainers, we selected the ”*nuclei*” pre-trained model). We also trained our own version using only the first frame of the movie, the manual segmentation of which is already cumbersome and takes approximately 6 hours using Napari [31] with a dedicated plugin. Examples of qualitative performances of all models can be observed on Figure 7b. Ground-truth-based metrics, shown on Figure 8a, are computed using all other frames. We noticed many FN errors due to objects cut at the frame borders in most models (particularly Stardist3D, Figure 7b). Since those errors are typically linked to implementation details rather than a fundamental limit of the network model, we sought to fairly compare the performances of Cellpose and Stardist3D using cropped predictions. For the centroid-based F1 score, we observe that a hypothetical ranking of the algorithms based on this metric would be independent from the choice of the threshold distance τ_c. Stardist3D performs better than Cellpose for pre-trained models, while the two reach similar values of $F1_c$ above their pre-trained counterparts once trained. Secondly, the IoU-based F1 score provides a similar overall ranking, but with much better performance for the pre-trained Stardist3D model compared to the pre-trained Cellpose one for the lower values of the IoU threshold. Trained models are again close to each other for high levels of the IoU threshold, but Cellpose lacks in performance for lower values of IoU. In order to evaluate our score on this dataset, we compute our heuristic on the search radius SR by visually inferring d_{err} and d_0 on a few slices.

We then analyze the performance of our score applied to this dataset. Results are shown on Figure 8b, where we present the average scores (averaged on each frame) obtained with each segmentation tool along with the average score obtained by applying our method on the ground truth segmentation. The score obtained by the ground-truth is smaller than the theoretically maximal value 1 due to ambiguities induced not

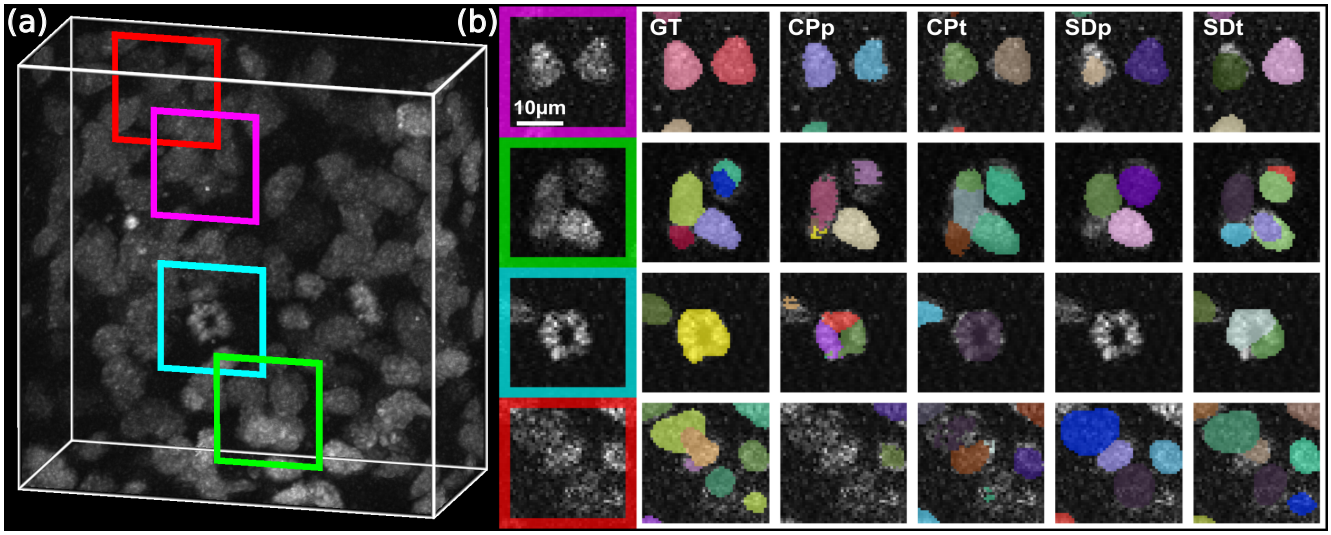

Fig. 7. Visual example of typical Stardist3D and Cellpose response on challenging 3D live organoid cells data. (a) Here, a single frame of the dataset is rendered in 3D using Napari. The dataset exhibits high cell density with heterogeneity in cellular dynamics and appearance. (b) Models tested (CP,SD: Cellpose, Stardist3D; p,t: pre-trained, trained) are compared to the ground-truth segmentation (GT) in four illustrative ROIs: two highly visible nuclei, a cluster of nuclei, a dividing nucleus, and a cluster with bad signal-to-noise ratio at the border of the field of view.

by errors in segmentation but by the dynamics itself (e.g. *bona fides* ambiguities due to large displacement combined with high density, or a real division event that would be interpreted as a split error). First, we observe correct overall ranking for the pre-trained versions of Cellpose and Stardist3D models. Our score predicts correctly the clear difference in performance measured in both $\mathrm{F1}_c$ and $\mathrm{F1}_{IoU}$. Second, the scores for the two trained models are close which reflect their proximity in centroid-based F1 score, but the overall ranking predicts that Stardist3D perform better, which is coherent with the difference in $\mathrm{F1}_{IoU}$.

Finally, we observe that the score associated with the pre-trained Stardist3D model reached the average score measured with the ground-truth. As such, it does not reflect the rank established by the true F1 score as opposed to Cellpose-pre-trained, Cellpose-trained and Stardist3D-trained. Further measurement and visual inspection performed to understand this discrepancy (Figure 8 and data not shown) suggest that Cellpose and Stardist3D exhibit similar rates but possibly different subtypes of false negatives. Indeed, on the one hand, cell clusters appear to be systematically missed by Cellpose but correctly detected by Stardist3D. On the other hand, isolated cells with unusual shapes seem to produce the reverse performance. The recent benchmarking study on both algorithms by Kleinberg and colleagues confirm that intuition [28]. As such, the discrepancy might stem not only from the mix of false negatives and false positives rates, but also from the cellular environment: whether it is in adherent cellular clusters or a freely moving cell in the aggregate. This could explain the overestimation of our quality score on the pre-trained version of Stardist3D since cell tracking in a relatively static cluster may present less variability in their trajectory-to-measurement association than freely moving cells. We discuss the implication of this hypothesis in our concluding Section.

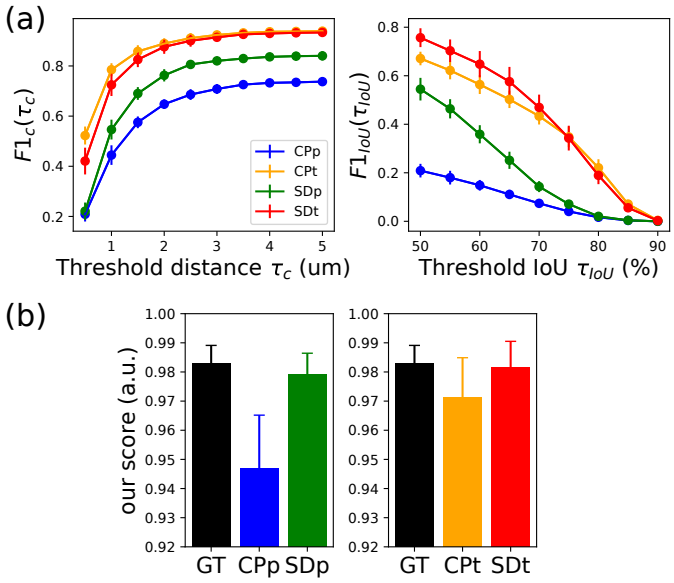

Fig. 8. Our evaluation score can accurately rank three of the four models studied on the organoid dataset. (a) We use the F1 scores based on centroid detection and on instance segmentation to rank the models using the ground-truth segmentation. We observe better performance from trained models. (b) Our score provides a similar ranking except for the pre-trained model of Stardist3D.

B. Carcinoma cells dataset

Here we show that our approach can be applied successfully to another dataset with different characteristics. The carcinoma cells dataset shows MDA231 human breast carcinoma cells infected with a pMSCV vector including the GFP sequence, embedded in a collagen matrix. For further informations, we refer the reader to the Cell Tracking Challenge [12]. This dataset was chosen as a complementary to the organoid cells dataset, as it is much sparser but with higher shape

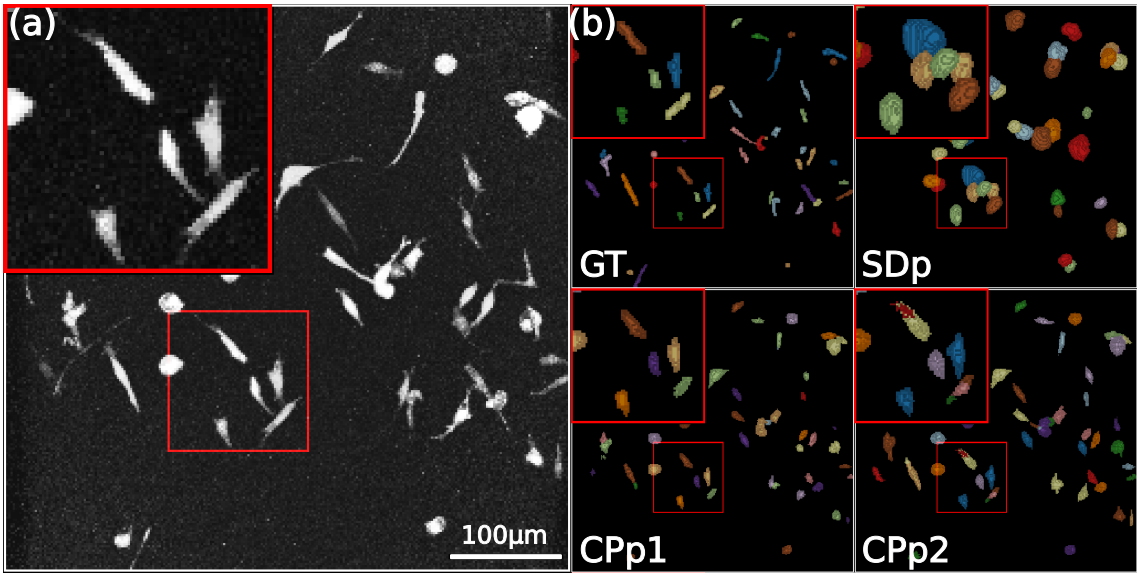

Fig. 9. Visual example of typical Stardist3D and Cellpose response on 3D live carcinoma cells data. (a) Here, a single frame of the dataset is rendered as a maximum intensity projection along the Z axis. The dataset exhibits sparse cell density but with high heterogeneity in cellular dynamics and appearance. (b) Models tested (SDp is the same Stardist3D pre-trained model as in Figure 7 and CPp1 and CPp2 are the "*cyto*" and "*cyto2*" pre-trained models of Cellpose) are compared to the ground-truth segmentation (GT). The inset displays an illustrative ROI.

heterogeneity. A frame is illustrated on Figure 9(a). We tested the pre-trained Stardist3D model and two pre-trained Cellpose models (models "*cyto*" and "*cyto2*"), which are compared to the ground truth on Figure 9(b). Notably, the Stardist3D model does not handle elongated cells well due to their non-star-convex shapes and also likely due to a lack of similar training data. Thus, it produces many splits and false positives errors, while the two Cellpose models seem to produce accurate segmentations.

The centroid-based F1 metric of Figure 9(a) reflects the much better performances of the Cellpose models compared to Stardist3D, but is not enough to separate the former two. From the IoU-based F1 score, we see that Cellpose model *cyto* has improved segmentation performances compared to model *cyto2*, which itself is marginally superior to Stardist3D. We note that despite visually identifying Cellpose models as accurate on Figure 9(b), all models have rather mediocre IoU-based F1 scores. This is easily explained by visually comparing the ground truth and Cellpose segmentations (Figure 10(b)), the ground truth being limited to the cell body and excluding protrusions while CellPose is typically more precise. Overall, the *cyto* model performs better than *cyto2*.

Finally, we show the performances given by our score on Figure 10(b). The search radius SR was still chosen using the previous heuristic and the score is again averaged on all frames. We see that the score accurately ranks the three detectors, with a relatively low value for Stardist3D compared to the two Cellpose models. The latter are also correctly separated, with *cyto* ranking higher than *cyto2*.

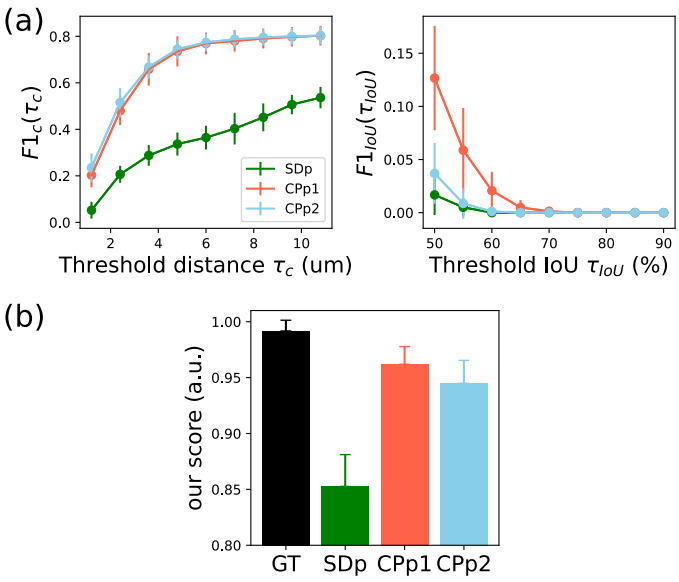

Fig. 10. Our evaluation score can accurately rank all models studied on the carcinoma cells dataset. (a) We use the F1 scores based on centroid detection and on instance segmentation to rank the models using the ground-truth segmentation. (b) Our score provides a similar ranking.

VII. Discussion

Our results demonstrate how using dynamical information measured through live cell experiments can be a powerful approach for the ranking of multiple detectors independently of their design. Indeed, our data shows that a score that merely described ambiguities in trajectory-to-detection assignment responds strongly to merging, splitting and false positive errors. Specifically, we show on simulations that our score approximates the F1 metric, a score commonly used for segmentation evaluation [9]. While dynamic ambiguities cannot detect individual FN, our score nevertheless responds well to false negatives in more realistic scenarios with mixed error types. In those mixed scenarios, the ranking between two detectors with different FN rates is preserved as long as the

rate of FP does not exceed 40%. In practice, a detector with a 40% FP rate is performing particularly poorly and can be easily dismissed through visual inspection. For comparison, the worst false positive rate we can measure across all models (some of them pre-trained on different datasets by other labs) using the centroid-based F1 is 15% on our challenging dataset. The scripts for the analysis of the score are available on Github at github.com/bioimage-mining-group/disco-wight, and we plan to make the code for the simulations freely available in the near future. Our score also shows very promising results when tested on experimental datasets with limited acquisition frequency and signal-to-noise ratio. We are indeed able to correctly rank the performances of deep-learning-based detectors, across pre-trained models and across trained models taken separately. However, we also show that trackability may be overestimated depending on the area of performance of a given detector. This indicates that a deeper analysis of the dynamics of segmentation candidates may help in improving ranking performances in those specific scenarios. As such, our work shows the potential for using dynamics to detect segmentation errors, but it also opens a variety of paths for future work and exciting perspectives for truly unsupervised validation of cell segmentation. The remainder of this discussion focuses on both of those aspects.

A first path of improvement is linked to sensitivity toward false negatives. This can be improved by injecting more *a priori* on the lifetime of our target, which is here virtually infinite in our organoid scenarios. This approach would help reduce the likelihood of the trajectory-to-death association and increase instability. A second challenge that we did not explore lies in the distinction between split events and cell divisions. This could be addressed using an *a priori* on instantaneous phase transition similar to previous works [20], [32]. Finally, we will further test the impact of the heterogeneity of the underlying biophysical processes and the sparse nature of mosaic imaging techniques This will require new methods for the simulation of cellular aggregates with a mixture of clusters and freely flowing cells as well as new imaging experiments imaged in non-mosaic conditions.

To conclude, our work paves the way toward the automated exploration of detection errors at the single cell level and the routine, annotation-free, evaluation of detection approaches that typically work as black boxes. An unsupervised score may also help in evaluating the need for retraining, further annotation, or even be part of the loss function used to train deep neural networks. It is our hope that those results will popularize the combination of live cell imaging and stochastic inferences to facilitate the validation of image-based measurements carried out at large scale.

REFERENCES

[1] B.-C. Chen, W. R. Legant, K. Wang, L. Shao, D. E. Milkie, M. W. Davidson, C. Janetopoulos, X. S. Wu, J. A. Hammer, Z. Liu, B. P. English, Y. Mimori-Kiyosue, D. P. Romero, A. T. Ritter, J. Lippincott-Schwartz, L. Fritz-Laylin, R. D. Mullins, D. M. Mitchell, J. N. Bembenek, A.-C. Reymann, R. Böhme, S. W. Grill, J. T. Wang, G. Seydoux, U. S. Tulu, D. P. Kiehart, and E. Betzig, "Lattice light-sheet microscopy: Imaging molecules to embryos at high spatiotemporal resolution," *Science*, vol. 346, no. 6208, p. 1257998.

[2] R. Galland, G. Grenci, A. Aravind, V. Viasnoff, V. Studer, and J.-B. Sibarita, "3D high- and super-resolution imaging using single-objective SPIM," *Nature Methods*, vol. 12, no. 7, pp. 641–644.

[3] K. M. Dean, P. Roudot, E. S. Welf, G. Danuser, and R. Fiolka, "Deconvolution-free Subcellular Imaging with Axially Swept Light Sheet Microscopy," *Biophysical Journal*, vol. 108, no. 12, pp. 2807–2815.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, vol. 9351, pp. 234–241, series Title: Lecture Notes in Computer Science.

[5] M. Weigert, U. Schmidt, R. Haase, K. Sugawara, and G. Myers, "Star-convex Polyhedra for 3D Object Detection and Segmentation in Microscopy," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Snowmass Village, CO, USA: IEEE, pp. 3655–3662.

[6] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, "Cellpose: a generalist algorithm for cellular segmentation," *Nature Methods*, vol. 18, no. 1, pp. 100–106.

[7] P. Roudot, W. R. Legant, Q. Zou, K. M. Dean, T. Isogai, E. S. Welf, A. F. David, D. W. Gerlich, R. Fiolka, E. Betzig, and G. Danuser, "u-track 3D: measuring and interrogating dense particle dynamics in three dimensions," Cell Biology, preprint.

[8] A. Beghin, A. Kechkar, C. Butler, F. Levet, M. Cabillic, O. Rossier, G. Giannone, R. Galland, D. Choquet, and J.-B. Sibarita, "Localization-based super-resolution imaging meets high-content screening," *Nature Methods*, vol. 14, no. 12, pp. 1184–1190.

[9] J. C. Caicedo, J. Roth, A. Goodman, T. Becker, K. W. Karhohs, M. Broisin, C. Molnar, C. McQuin, S. Singh, F. J. Theis, and A. E. Carpenter, "Evaluation of Deep Learning Strategies for Nucleus Segmentation in Fluorescence Images," *Cytometry Part A*, vol. 95, no. 9, pp. 952–965.

[10] B. Audelan and H. Delingette, "Unsupervised quality control of segmentations based on a smoothness and intensity probabilistic model," *Medical Image Analysis*, vol. 68, p. 101895.

[11] L. Beccari, N. Moris, M. Girgin, D. A. Turner, P. Baillie-Johnson, A.-C. Cossy, M. P. Lutolf, D. Duboule, and A. M. Arias, "Multi-axial self-organization properties of mouse embryonic stem cells into gastruloids," *Nature*, vol. 562, no. 7726, pp. 272–276.

[12] V. Ulman, M. Maska, K. E. G. Magnusson, O. Ronneberger, C. Haubold, N. Harder, P. Matula, P. Matula, D. Svoboda, M. Radojevic, I. Smal, K. Rohr, J. Jalden, H. M. Blau, O. Dzyubachyk, B. Lelieveldt, P. Xiao, Y. Li, S.-Y. Cho, A. C. Dufour, J.-C. Olivo-Marin, C. C. Reyes-Aldasoro, J. A. Solis-Lemus, R. Bensch, T. Brox, J. Stegmaier, R. Mikut, S. Wolf, F. A. Hamprecht, T. Esteves, P. Quelhas, O. Demirel, L. Malmstrom, F. Jug, P. Tomancak, E. Meijering, A. Munoz-Barrutia, M. Kozubek, and C. Ortiz-de Solorzano, "An objective comparison of cell-tracking algorithms," *Nature Methods*, vol. 14, no. 12, pp. 1141–1152.

[13] M. K. Driscoll and G. Danuser, "Quantifying Modes of 3D Cell Migration," *Trends in Cell Biology*, vol. 25, no. 12, pp. 749–759.

[14] M. Koziński, A. Mosinska, M. Salzmann, and P. Fua, "Tracing in 2D to reduce the annotation effort for 3D deep delineation of linear structures," *Medical Image Analysis*, vol. 60, p. 101590.

[15] U. Günther, K. I. S. Harrington, R. Dachselt, and I. F. Sbalzarini, "Bionic Tracking: Using Eye Tracking to Track Biological Cells in Virtual Reality," in *Computer Vision – ECCV 2020 Workshops*, A. Bartoli and A. Fusiello, Eds. Cham: Springer International Publishing, vol. 12535, pp. 280–297, series Title: Lecture Notes in Computer Science.

[16] H. Fehri, A. Gooya, Y. Lu, E. Meijering, S. A. Johnston, and A. F. Frangi, "Bayesian Polytrees With Learned Deep Features for Multi-Class Cell Segmentation," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3246–3260.

[17] D. Sarkar, G. Gompper, and J. Elgeti, "A minimal model for structure, dynamics, and tension of monolayered cell colonies," *Communications Physics*, vol. 4, no. 1, p. 36.

[18] K. Jaqaman, D. Loerke, M. Mettlen, H. Kuwata, S. Grinstein, S. L. Schmid, and G. Danuser, "Robust single-particle tracking in live-cell time-lapse sequences," *Nature Methods*, vol. 5, no. 8, pp. 695–702.

[19] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45.

[20] P. Roudot, L. Ding, K. Jaqaman, C. Kervrann, and G. Danuser, "Piecewise-Stationary Motion Modeling and Iterative Smoothing to Track Heterogeneous Particle Motions in Dense Environments," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5395–5410.

[21] S. Okuda, Y. Inoue, M. Eiraku, T. Adachi, and Y. Sasai, "Vertex dynamics simulations of viscosity-dependent deformation during tissue morphogenesis," *Biomechanics and Modeling in Mechanobiology*, vol. 14, no. 2, pp. 413–425.

[22] J. M. Osborne, A. G. Fletcher, J. M. Pitt-Francis, P. K. Maini, and D. J. Gavaghan, "Comparing individual-based approaches to modelling the self-organization of multicellular tissues," *PLOS Computational Biology*, vol. 13, no. 2, p. e1005387.

[23] P. Germann, M. Marin-Riera, and J. Sharpe, "ya||a: GPU-Powered Spheroid Models for Mesenchyme and Epithelium," *Cell Systems*, vol. 8, no. 3, pp. 261–266.e3.

[24] D. Pinheiro, R. Kardos, E. Hannezo, and C.-P. Heisenberg, "Morphogen gradient orchestrates pattern-preserving tissue morphogenesis via motility-driven unjamming," *Nature Physics*, vol. 18, no. 12, pp. 1482–1493.

[25] S. Tlili, J. Yin, J.-F. Rupprecht, M. A. Mendieta-Serrano, G. Weissbart, N. Verma, X. Teng, Y. Toyama, J. Prost, and T. E. Saunders, "Shaping the zebrafish myotome by intertissue friction and active stress," *Proceedings of the National Academy of Sciences*, vol. 116, no. 51, pp. 25 430–25 439.

[26] S. Kim, M. Pochitaloff, G. A. Stooke-Vaughan, and O. Campàs, "Embryonic tissues as active foams," *Nature Physics*, vol. 17, no. 7, pp. 859–866.

[27] K. M. Dean, P. Roudot, C. R. Reis, E. S. Welf, M. Mettlen, and R. Fiolka, "Diagonally Scanned Light-Sheet Microscopy for Fast Volumetric Imaging of Adherent Cells," *Biophysical Journal*, vol. 110, no. 6, pp. 1456–1465.

[28] G. Kleinberg, S. Wang, E. Comellas, J. R. Monaghan, and S. J. Shefelbine, "Usability of deep learning pipelines for 3D nuclei identification with Stardist and Cellpose," *Cells & Development*, vol. 172, p. 203806.

[29] "stardist/stardist-models." [Online]. Available: https://github.com/stardist/stardist-models

[30] "Models — cellpose 0.7.2 documentation." [Online]. Available: https://cellpose.readthedocs.io/en/latest/models.html

[31] N. Sofroniew, T. Lambert, K. Evans, J. Nunez-Iglesias, G. Bokota, P. Winston, G. Peña-Castellanos, K. Yamauchi, M. Bussonnier, D. Doncila Pop, A. Can Solak, Z. Liu, P. Wadhwa, A. Burt, G. Buckley, A. Sweet, L. Migas, V. Hilsenstein, L. Gaifas, J. Bragantini, J. Rodríguez-Guerra, H. Muñoz, J. Freeman, P. Boone, A. Lowe, C. Gohlke, L. Royer, A. PIERRÉ, H. Har-Gil, and A. McGovern, "napari: a multi-dimensional image viewer for Python."

[32] A. Genovesio, T. Liedl, V. Emiliani, W. Parak, M. Coppey-Moisan, and J.-C. Olivo-Marin, "Multiple particle tracking in 3-D+t microscopy: method and application to the tracking of endocytosed quantum dots," *IEEE Transactions on Image Processing*, vol. 15, no. 5, pp. 1062–1070.

Jules Vanaret completed his MS degree in physics from École Normale Supérieure of Lyon, France, and his engineering degree in computer science from École Centrale Lyon, France, in 2021. He is currently doing a Ph.D under the cosupervision of Frédéric Richard, Philippe Roudot and Sham Tlili in Marseille, France, within the Turing Center for Living System (CENTURI). He is interested in machine learning approaches to tackle physics problems.

Victoria Dupuis received her MS degree in molecular and cellular biology from Sorbonne University (Paris 6, France) in 2020. From 2020 to 2022, she worked as an engineer at the Marseille Developmental Biology Institute (AMU, France). Her research interests are developmental biology, organoids, and cell culture techniques.

Pierre-François Lenne studied physics at the University of Paris and École Normale Supérieure of Paris, France (1994), before completing his Ph.D. in soft matter physics at the University of Grenoble, France (1998). After postdoctoral research in the cell biology and biophysics unit of EMBL (Heidelberg, Germany), he joined the National Centre for Scientific Research (CNRS) at the Fresnel Institute in Marseille in 2000, where he developed and applied biophysical approaches to understand the dynamics of cell surfaces, using photonic structures in particular. In 2009, he was appointed group leader at the Institute for Developmental Biology of Marseille (IBDM) and research director at CNRS. With his group, he aims to understand the physical principles that underpin the morphogenesis of animals. He develops and applies quantitative approaches to observe, perturb and predict morphogenetic movements. He studies how cells generate and respond to mechanical forces, from supramolecular interactions at cell-cell contacts to the global shape of tissues.

Frédéric Richard received the MS degree in Mathematics and Vision from the Ecole Normale Supérieure of Cachan, and a Ph. D. in applied mathematics from the University Paris Descartes (actual University Sorbonne-Paris-Cité). For several years, he has been an Associate Professor at the University Paris Descartes, where he obtained the French habilitation in 2009. Since 2011, he is Professor of applied mathematics at Aix-Marseille University, member of the Institute of Mathematics of Marseille. His research interests are at the crossing of image processing, spatial statistics and data science. He designs and studies mathematical models and methods for the statistical analysis of structured data such as images, surfaces, or graphs. He applies his methods in different domains including medicine, neuroscience and industry.

Sham Tlili received a MS degree in Soft Matter and Complex Systems Physics at the International Centre for Fundamental Physics (ENS, Paris, France) in 2011, and her Ph.D. degree in Biophysics from the Paris Diderot University, France, in 2015. Her doctoral work hosted at the Matter and Complex Systems laboratory (Paris 7, France) and the Institute for Light and Matter (Lyon 1, France), focused on understanding both from experimental and theoretical aspects how cell-scale events contribute to tissue rheological properties, with an emphasis on in vitro tissues rheology (3D cell aggregates and cell monolayers). She then moved from 2015 to 2019 at the Mechanobiology Institute (NUS, Singapore) as a postdoctoral fellow, where she studied mechanical aspects of embryo morphogenesis (such as in the Zebrafish larva or the Drosophila embryo). Since 2020, she is a Research Investigator at the French National Centre for Scientific Research within the team "Physical approaches to cell dynamics and tissue morphogenesis" at the Marseille Developmental Biology Institute (AMU, France). Her main research focus is to understand the biophysical principles underlying embryonic organoids self-organization, by combining non-linear microscopy, 3D image analysis, tissue mechanical perturbations and physical modeling.

Philippe Roudot received the MS degree in computer science from the Institut National des Sciences Appliquées de Lyon, France, in 2010 and his Ph.D. degree in Signal Processing from the University of Rennes, France, in 2014. His doctoral work hosted at Inria (Rennes, France) and Curie Institute (Paris, France), focused on the unsupervised estimation of fluorescence lifetime in live cell microscopy through the study of the image formation model in frequency-domain fluorescence lifetime imaging. He has been a postdoctoral fellow from 2014 to 2020 at UT Southwestern Medical Center where he studies the quantification of macromolecular structures dynamics for the understanding of the cytoskeleton roles in cell morphogenesis. He is now a Group Leader within the Turing Center for Living System (CENTURI), where his group researches computer vision methods for the data driven study of physiological processes imaged across scales through fluorescence microscopy. Dr. Roudot is an appointed reviewer of the ISBI conference since 2014 and the biophysical journal since 2015. He was a Human Frontier Cross-Disciplinary Fellow from 2015 to 2018 and a Lindau Nobel Meeting alumni since 2016.