

# MPM: Joint Representation of Motion and Position Map for Cell Tracking

Junya Hayashida Kazuya Nishimura Ryoma Bise  
Kyushu University, Fukuoka, Japan, {bise@ait.kyushu-u.ac.jp}

## Abstract

Conventional cell tracking methods detect multiple cells in each frame (detection) and then associate the detection results in successive time-frames (association). Most cell tracking methods perform the association task independently from the detection task. However, there is no guarantee of preserving coherence between these tasks, and lack of coherence may adversely affect tracking performance. In this paper, we propose the Motion and Position Map (MPM) that jointly represents both detection and association for not only migration but also cell division. It guarantees coherence such that if a cell is detected, the corresponding motion flow can always be obtained. It is a simple but powerful method for multi-object tracking in dense environments. We compared the proposed method with current tracking methods under various conditions in real biological images and found that it outperformed the state-of-the-art (+5.2% improvement compared to the second-best).

## 1. Introduction

Cell behavior analysis is an important research topic in the fields of biology and medicine. Individual cells must be detected and tracked in populations to analyze cell behavior metrics including cell migration speed and frequency of cell division. However, it is too time consuming to manually track a large number of cells. Therefore, automatic cell tracking is required. Cell tracking in phase-contrast microscopy is a challenging multi-object tracking task. This is because the cells tend to have very similar appearance and their shapes severely deform. Therefore, it is difficult to identify the same cell among frames based on only its appearance. Moreover, cells in contact with each other often have blurry intercellular boundaries. In this case, it is difficult to identify touching cells from only one image. Finally, a cell may divide into two cells (cell mitosis), which is a very different situation from that of general object tracking.

The tracking-by-detection method is the most popular tracking paradigm because of the good quality of the detection algorithms that use convolutional neural networks (CNNs) for the detection step. These methods detect cells in each frame and then associate the detection results among

the time-frames by maximizing the sum of association scores of hypotheses, where the association step is often performed independently from the detection step [21, 31, 8]. They basically use hand-crafted association scores based on the proximity and similarity in shape of cells, but the scores only evaluate the similarity of individual cells; *i.e.*, they do not use the context of nearby cells. Moreover, the scores depend on the conditions of the time-lapse images, such as the frame interval, cell type, and cell density.

On the other hand, the positional relationship of nearby cells is important to identify the association when the density is high. Several methods have been proposed for using context information [32, 17]. Hayashida *et al.* [17] proposed a cell motion field (CMF) that represents the cell association between successive frames in order to directly predict cell motion (association) by using a CNN. This enables cell tracking at low frame rates. This method outperformed the previous methods that use hand-crafted association scores. However, it still independently performs the detection and association inference. This raises the possibility of incoherence between the two steps; *e.g.*, although a cell is detected, there is no response at the detected position in the CMF, and this would affect the tracking performance.

Many applications use multi-task learning [16, 19, 11, 45], *i.e.*, learning multiple-related tasks. This method basically has common layers express task-shared features and output branches that represent task-specific features. Multi-task learning improves the accuracy of the tasks compared with individual learning. However, it is not easy to preserve the coherence between the detection and association tasks in multi-task learning. If we simply apply a branch network, there is no guarantee that the coherence between these tasks will be preserved: *e.g.*, the network may detect a cell but not produce a corresponding association for the cell, which would adversely affect tracking performance.

Unlike such multi-task learning methods, we propose the Motion and Position Map (MPM) that jointly represents both detection and association for not only migration but also cell division, as shown in Fig. 1, where the distribution of magnitudes of MPM indicates the cell-position likelihood map at frame  $t$ ; the direction of the 3D vector encoded on a pixel in MPM indicates the motion direction

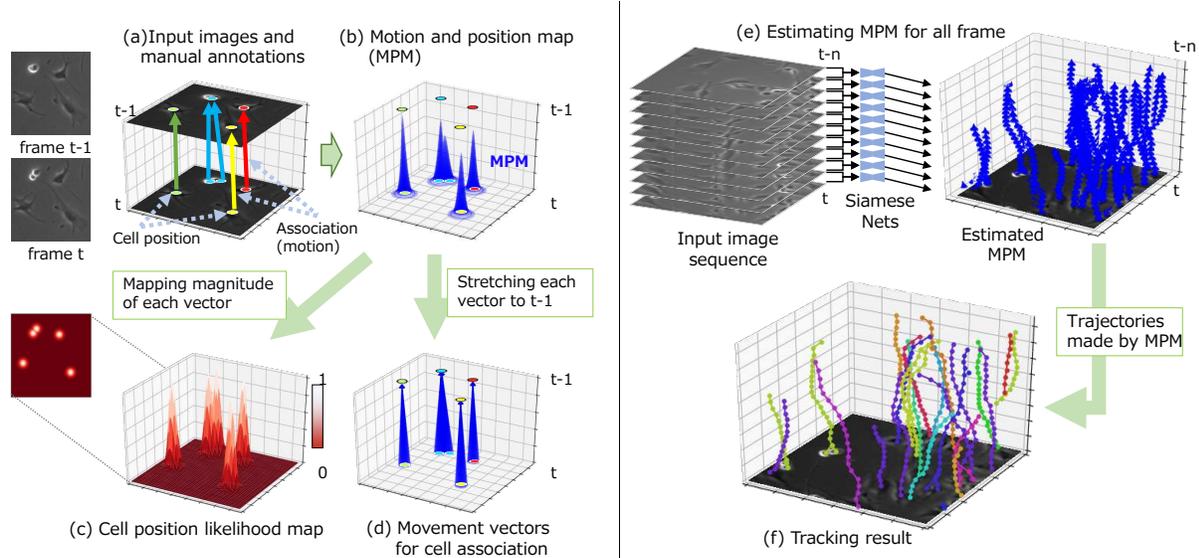


Figure 1. Overview of the proposed method. (a) Input images in two successive frames and manual annotations of cell positions in each frame and their association. (b) Motion and position map (MPM) that jointly represents both detection and association. MPM encodes the following two pieces of information: (c) a cell-position likelihood map, which is a magnitude map of MPM. The peaks of this map indicate the cells' positions; (d) motion vector for a cell from  $t$  to  $t - 1$ . (e) Estimating the MPM for all frames by using siamese MPM-Nets. (f) The overall tracking results when using MPM.

from the pixel at  $t$  to the cell centroid at  $t - 1$ . Since MPM represents both detection and association, it guarantees coherence; *i.e.*, if a cell is detected, an association of the cell is always produced. In addition, we can obtain the cell positions and their association for all frames by using siamese MPM networks, as shown in Fig. 1(e). Accordingly, even a very simple tracking method can be used to construct overall cell trajectories with high accuracy (Fig. 1(f)). In the experiments, our method outperformed the state-of-the-arts under various conditions in real biological images.

Our main contributions are summarized as follows:

- We propose the motion and position map (MPM) that jointly represents detection and association for not only migration but also cell division. It guarantees coherence such that if a cell is detected, the corresponding motion flow can always be obtained. It is a simple but powerful method for multi-object tracking in dense environments.
- By applying MPM-Net to the entire image sequence, we can obtain the detection and association information for the entire sequence. By using the estimation results, we can easily obtain the overall cell trajectories without any complex process.

## 2. Related work

**Cell tracking:** Cell tracking methods can be roughly classified into two groups: those with model-based evolution and those based on detection-and-association. The first

group uses sequential Bayesian filters such as particle filters [30, 38], or active contour models [24, 40, 43, 46]. They update a probability map or energy function of cell regions by using the tracking results of the previous frame as the initial state for the updates. These methods work well if cells are sparsely distributed. However, these methods may get confused when cells are close together with blurry inter-cellular boundaries or move over a long distance.

The second group of cell-tracking methods are based on detection and association: first they detect cells in each frame; then they determine associations between successive frames. Many cell detection/segmentation methods have been proposed to detect individual cells, by using level set [25], watershed [28], graph-cut [4], optimization using intensity features [7], physics-based-restoration [44, 39, 23], weakly-supervised [29] and CNN-based detection methods [33, 34, 1, 27, 3]. Moreover, many optimization methods have been proposed to associate the detected cells. They maximize an association score function, including linear programming for frame-by-frame association [21, 5, 6, 46, 42] and graph-based optimization for global data association [8, 36, 39, 15]. These methods basically use the proximity and shape similarity of cells for making hand-crafted association scores. These association scores depend on the cell culture conditions, such as the frame interval, cell types and cell density. In addition, the scores only evaluate the similarity of individual cells, not the context of nearby cells. On the other hand, a human expert empirically uses the positional relationship of nearby

cells in dense cases. Several methods have been proposed on how to use such context information [32, 17, 33]. Payer *et al.* [32] proposed a recurrent stacked hourglass network (ConvGRU) that not only extracts local features, but also memorizes inter-frame information. However, this method requires annotations for all cell regions as training data, and it does not perform well when the cells are small. Hayashida *et al.* [17] proposed a cell motion field that represents the cell association between successive frames that can be estimated with a CNN. Their method is a point-level tracking, and thus it dramatically reduces annotation costs compared with pixel-level cell boundary annotation. These methods have been shown to outperform those that use hand-crafted association scores. However, the method still performs the detection and association inference independently. This causes incoherence between steps, which in turn affects tracking performance.

**Object tracking for general object:** For single-object tracking that finds the bounding box corresponding to given initial template, Siamese networks have been often used to estimate the similarity of semantic features between the template and target images [20, 22]. To obtain a good similarity map, this approach is based on representation learning. However, in the cell tracking problem, there may be many cells that have similar appearances; thus, the similarity map would have a low confidence level. Tracking-by-detection (detection-and-association) is also a standard approach in multi-object tracking (MOT) for general objects. In the association step, features of objects and motion predictions [26, 12] are used to compute a similarity/distance score between pairs of detection and/or tracklets. Re-identification is often used for feature representation [14, 2]. This method does not use the global spatial context, and the detection and association processes are separated. Optimization techniques [41, 35] have been proposed for jointly performing the detection and association processes; they jointly find the optimal solution of detection and association from a candidate set of detections and associations. Sun *et al.* [37] proposed the Deep Affinity Network (DAN). Although this method jointly learns a feature representation for identification and association, the detection step is independent of these learnings.

**Multi-task learning:** Multi task learning (MTL) has been widely used in computer vision for learning similar tasks such as pose estimation and action recognition [16]; depth prediction and semantic classification [19, 11]; and for room floor plans [45]. Most multi-task learning network architectures for computer vision are based on existing CNN architectures [19, 9, 18]. These architectures basically have common layers and output branches that express both task-shared and task-specific features. For example, cross-stitch networks [19] contain one standard CNN per task, with cross-stitch units to allow features to be shared across tasks.

Such architectures require a large number of network parameters and scale linearly with the number of tasks. In addition, there is no guarantee of coherence among the tasks if we simply apply the network architecture to a multi-task problem such as detection and association. This means that we need to design a new loss function or architecture to preserve the coherence depending on the pairs of the tasks. To the best of our knowledge, multi-task learning methods specific to our problem have not been proposed.

Unlike these methods, we propose MPM for jointly representing the detection and association tasks at once. The MPM can be estimated using simple U-net architecture, and it guarantees coherence such that if a cell is detected, the corresponding motion flow can always be obtained.

### 3. Motion and Position Map (MPM)

Our tracking method estimates the Motion and Position Map (MPM) with a CNN. The MPM jointly represents the position and moving direction of each cell between successive frames by storing a 3D vector on a 2D plane. From the distribution of the 3D vectors in the map, we can obtain the positions of the cells and their motion information. First, the distribution of the magnitudes of the vectors in the MPM shows the likelihoods of the cell positions in frame  $t$ . Second, the direction of the 3D vector on a cell indicates the motion direction from the pixel at  $t$  to the cell centroid at  $t - 1$ . Note that we treat the cell motion from  $t$  to  $t - 1$  in order to naturally define a cell division event when a mother cell divides into two daughter cells by one motion from a cell. By using MPM as the ground-truth, it is possible to use a simple CNN model that has a U-net architecture, that we call MPM-Net, to perform the simultaneous cell detection and association tasks.

First, we explain how to generate the ground truth of MPM from the manual annotation. Fig. 2 shows the simple example of the annotations for the motion of a single cell and its MPM. The manual annotation for an image sequence expresses the position of each cell position in terms of 3D orthogonal coordinates by adding a frame index to the annotated 2D coordinate, where the same cell has the same ID from frame to frame. We denote the annotated cell positions for cell  $i$  at frame  $t - 1$  and  $t$  as  $\mathbf{a}_{t-1}^i, \mathbf{a}_t^i$  respectively. They are defined as:

$$\begin{cases} \mathbf{a}_{t-1}^i &= (x_{t-1}^i, y_{t-1}^i, t-1)^T, \\ \mathbf{a}_t^i &= (x_t^i, y_t^i, t)^T. \end{cases} \quad (1)$$

Although the annotated position for a cell is a single point at  $t$ , manual annotation may not always correspond to the true position. Similar to [17], we make a likelihood map of cell positions, where an annotated cell position becomes a peak and the value gradually decreases in accordance with a Gaussian distribution, as shown in Fig. 2. Similarly, the

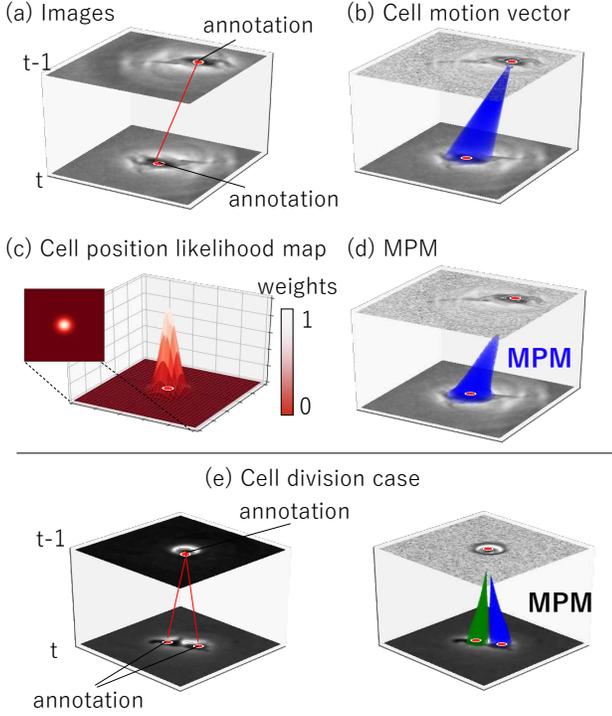


Figure 2. Example of the generated MPM from the given annotated cell positions at  $t$  and  $t - 1$ . Note that this example shows the enlarged images for a single cell for explanation. (a) Input images with the annotation points. The red points indicate the annotated points on the rough centroids of cells. (b) Motion vectors from the neighbor points around the annotated point at  $t$  to the annotated point  $t - 1$ . (c) Cell position likelihood map that indicates the weight map of the vector of MPM. (d) MPM that represents both the motion vector and the position likelihood map, where it encodes the 3D vector at each pixel and the magnitude of the vector indicates the cell position likelihood, the direction of the vector indicates the direction of the motion. (e) Example of the cell division case, where two cells are associated to the single cell.

motion vectors are represented as the 3D vectors pointing to  $\mathbf{a}_{t-1}^i$  from the nearby pixels of  $\mathbf{a}_t^i$ .

To generate ground truth MPM  $\mathbf{C}_{t-1,t}^i$ , we first generate individual MPM  $\mathbf{C}_{t-1,t}^i$  for each cell  $i$  in frame  $t - 1$  and  $t$ . Each individual MPM is obtained by storing 3D vectors pointing to  $\mathbf{a}_{t-1}^i$  from the neighboring region of  $\mathbf{a}_t^i$ . Let  $\mathbf{v}_{t-1,t}^i = \mathbf{a}_{t-1}^i - \mathbf{p}_t$  be the motion vector from each coordinate  $\mathbf{p}_t = (x_t, y_t, t)^T$  at frame  $t$  to the annotated position  $\mathbf{a}_{t-1}^i$ .  $\mathbf{C}_{t-1,t}^i$  is defined as:

$$\mathbf{C}_{t-1,t}^i(\mathbf{p}_t, \mathbf{v}_{t-1,t}^i) = w(\mathbf{p}_t) \frac{\mathbf{v}_{t-1,t}^i}{\|\mathbf{v}_{t-1,t}^i\|_2}, \quad (2)$$

where  $w(\cdot)$  is a weight function that takes each coordinate at frame  $t$  as input and returns a scalar value. Since the motion vector  $\mathbf{v}_{t-1,t}^i / \|\mathbf{v}_{t-1,t}^i\|_2$  is a unit vector in Eq.(2),  $w(\cdot)$

is a function that represents the magnitude of each vector in MPM.  $w(\cdot)$  is a Gaussian function with the annotation coordinate  $\mathbf{a}_t^i$  as the peak, and is defined as:

$$w(\mathbf{p}_t^i) = \exp\left(-\frac{\|\mathbf{a}_t^i - \mathbf{p}_t\|_2^2}{\sigma^2}\right), \quad (3)$$

where  $\sigma$  is a hyper-parameter that indicates the standard deviation of the distribution and controls the spread of the peak.

The vector stored at each coordinate of MPM  $\mathbf{C}_{t-1,t}^i$  is the vector of the individual MPMs with the largest value of the coordinates among all the magnitudes of the individual MPMs. The MPM  $\mathbf{C}_{t-1,t}$  that aggregates all the individual MPM is defined as:

$$\mathbf{C}_{t-1,t}(\mathbf{p}_t) = \mathbf{C}_{t-1,t}^{\hat{i}}(\mathbf{p}_t), \quad (4)$$

$$\hat{i} = \arg \max_i \|\mathbf{C}_{t-1,t}^i(\mathbf{p}_t)\|. \quad (5)$$

Fig. 2(d) shows an example of the MPM generated based on the annotation (a). One of the advantage of the MPM is that can represent the cell division case that two daughter cells are associated to a mother cell, as shown in Fig. 2(e).

A U-Net is trained with the ground truth of the MPM by using a loss function that is the mean squared error (MSE) between the training data and the estimated data, where two images are treated as two channels of the input, and the outputs of a pixel are 3 channels of the 3D vector. Let the ground-truth of MPM be  $\mathbf{c}$  and the estimated MPM be  $\hat{\mathbf{c}}$ . The loss function is defined as:

$$\text{Loss}(\mathbf{c}, \hat{\mathbf{c}}) = \frac{1}{n} \sum_{i=1}^n (\|\mathbf{c}_i - \hat{\mathbf{c}}_i\|_2^2 + (\|\mathbf{c}_i\|_2 - \|\hat{\mathbf{c}}_i\|_2)^2) \quad (6)$$

where the first term  $\|\mathbf{c}_i - \hat{\mathbf{c}}_i\|_2^2$  is a square error between  $\mathbf{c}$  and  $\hat{\mathbf{c}}$ , and the second term  $(\|\mathbf{c}_i\|_2 - \|\hat{\mathbf{c}}_i\|_2)^2$  is a square error between the magnitudes of  $\mathbf{c}$  and  $\hat{\mathbf{c}}$ . Here, we empirically added the second term that directly reflects the error of the magnitude (i.e., detection). Even though we use only the first term, it works, but the proposed loss function was stable.

## 4. Cell tracking by MPM

Fig. 3 shows an overview of cell detection and association in inference. For frame  $t$  and  $t - 1$  (Fig. 3(a)), we estimate the MPM  $\mathbf{C}_{t-1,t}$  by using the trained MPM-Net (Fig. 3(b)). The positions of cells in frame  $t$  are detected from the magnitude of  $\mathbf{C}_{t-1,t}$  (Fig. 3(c)(d)), and the cells in frame  $t - 1$  are associated with each 3D vector stored at the detected positions (Fig. 3(e)).

### 4.1. Cell detection by the magnitude of MPM

The cell-position likelihood map  $\mathbf{L}_{t-1,t}$  at frame  $t$  is obtained from the distribution of magnitudes of MPM  $\mathbf{C}_{t-1,t}$ .

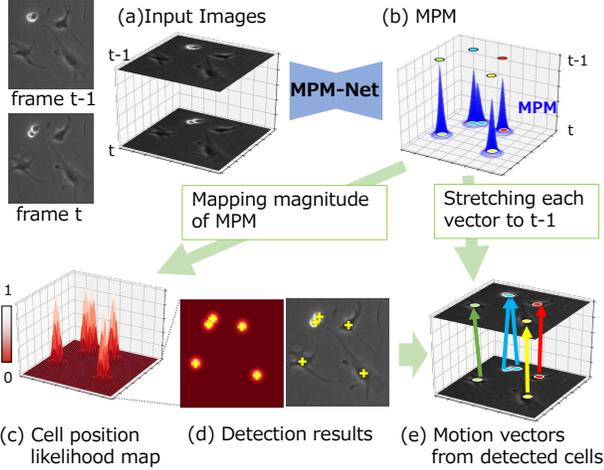


Figure 3. Overview of cell detection and association in inference. (a) Input images. (b) MPM estimated by MPM-Net. (c) Cell-position likelihood map determined from the magnitude of MPM. (d) The peaks are detected as cells at  $t$ . (e) The cell motion from  $t$  to  $t - 1$  is estimated by the direction vector at the detected peak point of MPM.

The detected cell positions are the coordinates of the each local maximum of  $L_{t-1,t}$ . In the implementation, we applied Gaussian smoothing as a pre-processing against noise to  $L_{t-1,t}$ .

## 4.2. Cell association by MPM

The direction of each 3D vector of MPM  $C_{t-1,t}$  indicates the direction from position of the cell at  $t$  to position of the same cell at  $t - 1$ . The position at frame  $t - 1$  is estimated using the triangle ratio from each vector of  $C_{t-1,t}$  stored at each detected position in frame  $t$ , as shown in Fig. 4. Let the detected position of each cell  $i$  be  $\mathbf{m}_t^i = (x_t^i, y_t^i, t)^\top$ , and the 3D vector stored in  $\mathbf{m}_t^i$  be  $\mathbf{v}_{t-1,t}^i = (\Delta x, \Delta y, \Delta t)^\top$ . The estimated position  $\hat{\mathbf{m}}_{t-1}^i$  for each cell  $i$  at frame  $t - 1$  is defined as:

$$\hat{\mathbf{m}}_{t-1}^i = (x_t^i + \frac{\Delta x}{\Delta t}, y_t^i + \frac{\Delta y}{\Delta t}, t - 1)^\top. \quad (7)$$

We associate the detected cells in frame  $t$  with the tracked cells in frame  $t - 1$  on the basis of  $\hat{\mathbf{m}}_{t-1}^i$  and the previous confidence map  $L_{t-2,t-1}$  used for the cell detection in frame  $t - 1$ , as shown in Fig. 5. The estimated position  $\hat{\mathbf{m}}_{t-1}^i$  of cell  $i$  at  $t - 1$  may be different from the tracked point at  $t - 1$  which is located at a local maximum of cell-position likelihood map. Thus, we update the estimated position  $\hat{\mathbf{m}}_{t-1}^i$  to a certain local maximum that is the position of a tracked cell  $\mathbf{T}_{t-1}^j$  by using the hill-climbing algorithm to find the peak (the position of the tracked cell). Here, the  $j$ -th cell trajectory from  $s_j$  until  $t - 1$  is denoted as  $\mathbf{T}^j = \{\mathbf{T}_{s_j}^j, \dots, \mathbf{T}_{t-1}^j\}$ , where  $s_j$  indicates the frame in

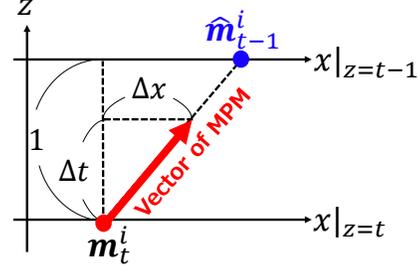


Figure 4. Estimation of the cell position at frame  $t - 1$

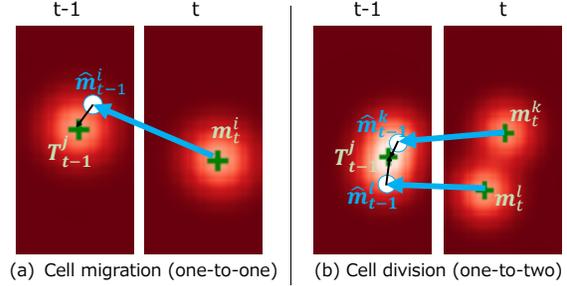


Figure 5. Illustration of association process. (a) The case of cell migration (one-to-one matching): the estimated position  $\hat{\mathbf{m}}_{t-1}^i$  is associated with the nearest tracked cell  $\mathbf{T}_{t-1}^j$ . (b) The case of cell division (one-to-two matching): two cells  $\mathbf{m}_{t-1}^k, \mathbf{m}_{t-1}^l$  are associated as daughter cells of  $\mathbf{T}_{t-1}^j$ .

which the  $j$ -th cell trajectory started. If  $\hat{\mathbf{m}}_{t-1}^i$  was updated to a position with  $\mathbf{T}_{t-1}^j$ , we associate  $\mathbf{m}_t^i$  with  $\mathbf{T}_{t-1}^j$  as the tracked cell  $\mathbf{T}_t^j = \mathbf{m}_t^i$ . If the confidence value at the estimated position at  $t - 1$  is zero (*i.e.*, the point is not associated with any cell),  $\mathbf{m}_{t-1}^i$  is registered as a newly appearing cell  $\mathbf{T}^{N_{t-1}+1}$ , where  $N_{t-1}$  is the number of cell trajectories until time  $t$  before the update.

If two estimated cell positions  $\mathbf{m}_{t-1}^k, \mathbf{m}_{t-1}^l$  were updated to one tracked cell position  $\mathbf{T}_{t-1}^j$ , these detected cells are registered as new born cells  $\mathbf{T}^{N+1} = \{\mathbf{m}_{t-1}^k\}, \mathbf{T}^{N+2} = \{\mathbf{m}_{t-1}^l\}$ , *i.e.*, daughter cells of  $\mathbf{T}^j$ , and the division information is also registered  $\mathbf{T}^j \rightarrow \mathbf{T}^{N_{t-1}+1}, \mathbf{T}^{N_{t-1}+2}$ . The cell tracking method using MPM can be performed very simply for both cell migration and division.

## 4.3. Interpolation for undetected cells

The estimated MPM at  $t$  may have a few false negatives because the estimation performance is not perfect. In this case, the trajectory  $\mathbf{T}^j$  is not associated with any cell and the frame-by-frame tracking for the cell is terminated at  $t - 1$  for the moment. The tracklet  $\mathbf{T}^j$  is registered as a temporarily track-terminated cell. During the tracking process using MPM, if there are newly detected cells including daughter cells from cell division at  $t - 1 + q$ , the method tries to

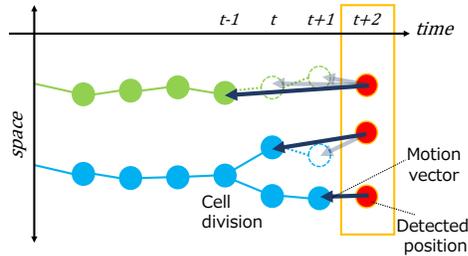


Figure 6. Illustration of interpolation process for undetected cells. The newly detected cells (red) are associated with the track-terminated cells (green) using MPM.

associate the temporarily track-terminated cell using MPM again, where the MPM is estimated by inputting the images at  $t - 1$  and  $t - 1 + q$  (*i.e.*, the time interval of the images is  $q$ ). Next, the cell position is estimated and updated by using frame-by-frame association. If the newly detected cell is associated with one of the temporarily track-terminated cells, the cell is then associated with the cell and the positions in the interval of frames, in which corresponding cells were not detected, are interpolated based on the vector of the MPM. If there are multiple temporarily track-terminated cells, this process is iterated on all track-terminated cells in ascending order of time. The method excludes a cell from the list of the track-terminated cells when the interval from the endpoint of the track-terminated cell to the current frame is larger than a threshold. Here, we should note that cells at  $t$  tend to be associated with another cell as cell divisions when a false negative occurs at  $t - 1$ . To avoid such false-positives of cell-division detection, the method tries to associate such daughter cells with the track-terminated cells. This cell tracking algorithm can be performed online.

Fig. 6 shows an example of this process, in which the tracking of the green trajectory is terminated at  $t - 1$  due to false negatives at  $t$ ,  $t + 1$ , and the red cell is detected at  $t + 2$ . In this case, the MPM is estimated for three frame intervals (the inputs are the images at  $t - 1$  and  $t + 2$ ), and it is associated with the track-terminated cell. Then, the cell positions at  $t$  and  $t + 1$  are interpolated using the vector of MPM and the tracking starts again.

## 5. Experiments

### 5.1. Data set

In the experiments, we used the open dataset [13] that includes time-lapse image sequences captured by phase-contrast microscopy since this is the same setup with our target to track. In this data-set, the point-level ground truth is given, in which the rough cell centroid was annotated with cell ID, and hundreds of mouse myoblast stem cells were cultured under four cell culture conditions that are different growth-factor mediums; a) Control (no growth factor), b) FGF2, c) BMP2, d) BMP2+FGF2, where there

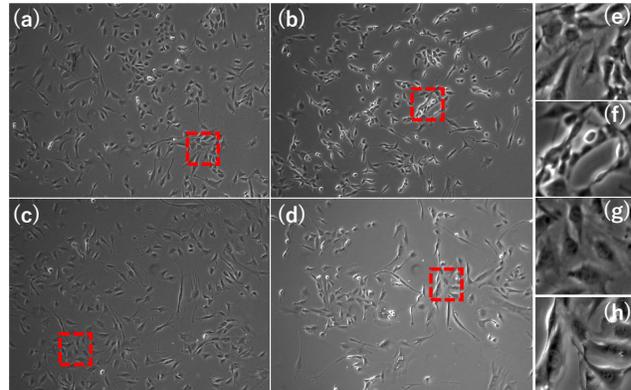


Figure 7. Example images from four conditions. a) Control, b) FGF2, c) BMP2, d) BMP2+FGF2, (e)-(h) the enlarged images of the red box in (a)-(d), respectively. The appearance of cells are different depending on the culture conditions. Cells often shrink and partially overlapped in FGF2, cells tend to be spread under BMP2, there are both cells who are spread or shrunk under BMP2+FGF2.

are four image sequences for each condition (total: 16 sequences). Since a stem cell is differentiated depending on the growth factor, the cell appearance changes over the four culture conditions as shown in Fig. 7.

The images were captured every 5 minutes and each sequence consists of 780 images with the resolution of  $1392 \times 1040$  pixels. All cells were annotated in a sequence in BMP2, and three cells were randomly picked up at the beginning of the sequence, and then the three cell's family trees through the 780-th image were annotated for all the other 15 sequences, where the number of the annotated cells increased with time due to cell division. The total number of annotated cells in the 16 sequences is 135859 [21].

Since our method requires the training data that all cells should be fully annotated in each frame, we additionally annotated cells in a part of frames under three conditions; Control: 100 frames, BMP2: 100 frames, FGF2+BMP2: 200 frames. Totally the annotated 400 frames were used as the training data, and the rest of the data was used as ground truth in the test. To show the robustness for the cross-domain data (the different culture conditions from training data) in the test, we did not make the annotation for FGF2. In all experiments, we set the interval value  $q$  for interpolation as 5.

### 5.2. Performance of cell Detection

The detection performance of our method compared with the three other methods including the state-of-the-art; Bensch [4] that is based-on graph-cut, Bise [8] that is based on physics model of phase-contrast microscopy [44], and Hayashida [17] (the state-of-the-art for this data-set) that is based on deep learning that estimates the cell-position likelihood map that is similar to ours. Here, the annotation of this dataset is the point-level annotation, and thus we could

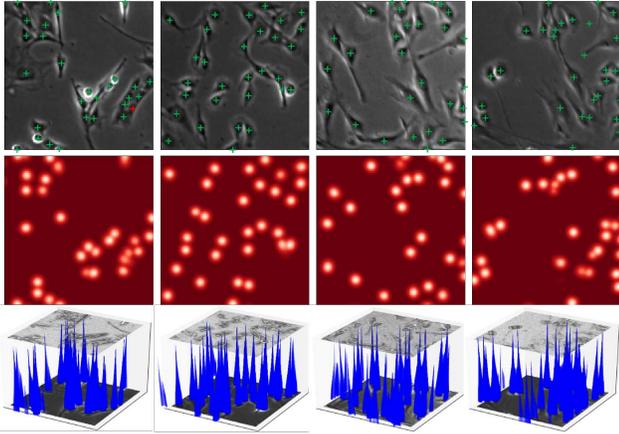


Figure 8. Examples of our cell detection results. Top: detection results. Green '+' indicates the true positive, red '+' indicates the false negative. Middle: cell position likelihood map generated from the estimated MPM (Bottom).

not apply the learning-based segmentation methods that require the pixel-level annotation for the individual cell regions. We used the fully-annotated sequence to measure the precision, recall, and F1-score as the detection performance metrics. For the training and tuning the parameters, we used the same training sequences.

Fig. 8 shows examples of our cell detection results. Our method successfully detected the cells that have various appearances, including spread cells, small brighter cells, and touching cells. Table 1 shows the results of the cell detection evaluation. The deep learning-based methods (Hayashida’s and ours) outperformed the other two methods. The detection method in [17] estimated the cell-position likelihood map by U-net, in which their method represents the detection but not for the association. The terms of recall and F1-score of our method are slightly better (+1.4%) than those of their method. Since our method uses the context from two successive frames in contrast to their method that detects cells each frame independently, we consider that it would be possible to improve the performance of detection.

### 5.3. Performance of cell Tracking

Next, we evaluated the tracking performance of our method with four other methods; Bensch [4] based-on frame-by-frame association with graph-cut segmentation; Chalfoun [10] based-on optimization for the frame-by-frame association with the segmentation results by [44]; Bise[8] based on spatial-temporal global data association; Hayashida [17] based on motion flow estimation by CNN. Hayashida’s method is the most related to the proposed method. Their method estimates the cell position map and motion flow by CNN, independently.

We used two quantitative criteria to assess the per-

Table 1. Cell detection performance on the terms of precision, recall and F1-score.

Method	<i>Bensch</i> [4]	<i>Bise</i> [8]	<i>Hayashida</i> [17]	<i>Ours</i>
precision	0.583	0.850	<b>0.968</b>	0.964
recall	0.623	0.811	0.902	<b>0.932</b>
F1-score	0.602	0.830	0.934	<b>0.948</b>

Table 2. Association Accuracy.

Method	Cont.	FGF2	BMP2	FGF2+ BMP2	Ave.
<i>Bensch</i> [4]	0.604	0.499	0.801	0.689	0.648
<i>Chalfoun</i> [10]	0.762	0.650	0.769	0.833	0.753
<i>Bise</i> [8]	0.826	0.775	0.855	0.942	0.843
<i>Hayashida</i> [17]	0.866	0.884	0.958	0.941	0.912
<i>Ours (MPM)</i>	<b>0.947</b>	<b>0.952</b>	<b>0.991</b>	<b>0.987</b>	<b>0.969</b>

Table 3. Target Effectiveness.

Method	Cont.	FGF2	BMP2	FGF2+ BMP2	Ave.
<i>Bensch</i> [4]	0.543	0.448	0.621	0.465	0.519
<i>Chalfoun</i> [10]	0.683	0.604	0.691	0.587	0.641
* <i>Li</i> [23]	0.700	0.570	0.630	0.710	0.653
* <i>Kanade</i> [21]	<b>0.830</b>	0.640	0.800	0.790	0.765
<i>Bise</i> [8]	0.733	0.710	0.788	0.633	0.771
<i>Hayashida</i> [17]	0.756	0.761	0.939	0.841	0.822
<i>Ours (MPM)</i>	0.803	<b>0.829</b>	<b>0.958</b>	<b>0.911</b>	<b>0.875</b>

formance: association accuracy and target effectiveness [21]. To compute association accuracy, each target (human-annotated) was assigned to a track (computer-generated) for each frame. The association accuracy was computed as the number of true positive associations divided by the number of associations in the ground-truth. If a switch error occurred between two cells A and B, we count the two false positives (A → B, B → A). To compute target effectiveness, we first assign each target to a track that contains the most observations from that ground-truth. Then target effectiveness is computed as the number of the assigned track observations over the total number of frames of the target. It indicates how many frames of targets are followed by computer-generated tracks. This metric is a stricter metric. If a switching error occurs in the middle of the trajectory, the target effectiveness is 0.5.

Table 2 and 3 show the performance of the association accuracy and the target effectiveness, respectively. In these tables, the average scores for each condition (three or four sequences in each condition) are denoted <sup>1</sup>. In Table 3, we show the additional two methods as the reference; Li [23]

<sup>1</sup>The scores of the target effectiveness of \*Li and \*Kanade were evaluated by the same data-set in their papers. Since the association accuracy of their methods [23, 21] were not described in their papers, we could not add their performance at the Table 2.

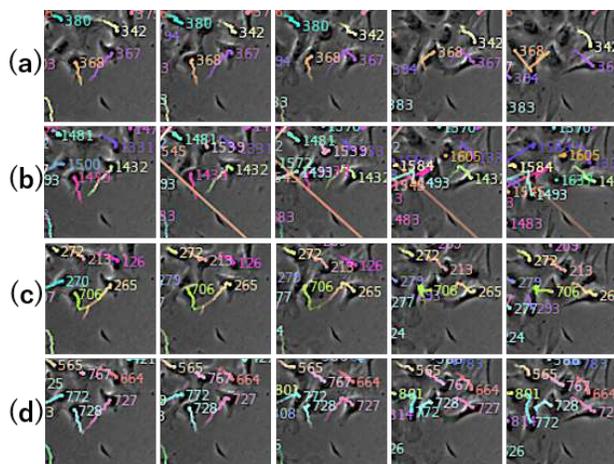


Figure 9. Tracking results from each compared method under BMP2. (a) Bise, (b) Chalfoun, (c) Hayashida, (d) ours. Horizontal axis indicate the time.

based on level-set tracking and Kande [21] based on frame-by-frame optimization. The performances of Bensch, Chalfoun, Li were sensitive depending on the culture conditions, and the performances were not good on both the metric terms, in particular, (FGF2) and (BMP2+FGF2). We consider that the sensitivity of their detection methods adversely affects the tracking performance. The methods of Bise and Kanade that use the same detection method [44] achieved better accuracy compared to the three methods thanks to the better detection. The state-of-the-art method (Hayashida) furthermore improved the performances (Fig. 9(c)) compared with the other current methods by estimating cell motion flow and position map independently. Our method (MPM) outperformed all the other methods (Fig. 9(d)). It improved +5.7% of association accuracy and +5.1% of the target-effectiveness on the average score compared to the second-best (Hayashida).

Fig. 10 shows the tracking results from ours under each condition, in which our method correctly tracked the various appearance of cells. Fig. 11 shows the tracking results in the case when a cell divided into two cells in the enlarged image-sequence. Our method could correctly identify the cell division and track all cells in the region.

## 6. Conclusion

In this paper, we proposed the motion and position map (MPM) that jointly represents both detection and association, which can be represented in the cell division case not only migration. The MPM encodes a 3D vector into each pixel, where the distribution of magnitudes of the vectors indicates the cell-position likelihood map, and the direction of the vector indicates the motion direction from the pixel at  $t$  to the cell centroid at  $t - 1$ . Since MPM represents both detection and association, it guarantees coher-

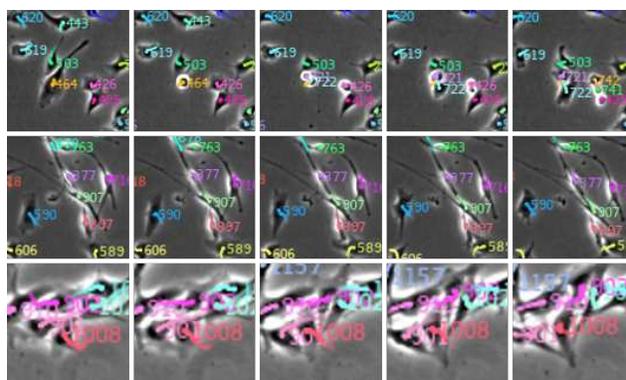


Figure 10. Tracking results from our method under each culture conditions; (a) Control, (b) FGF2, (c) BMP2+FGF2. Although the cell appearances are different depending on the conditions, our method correctly tracked the cells.

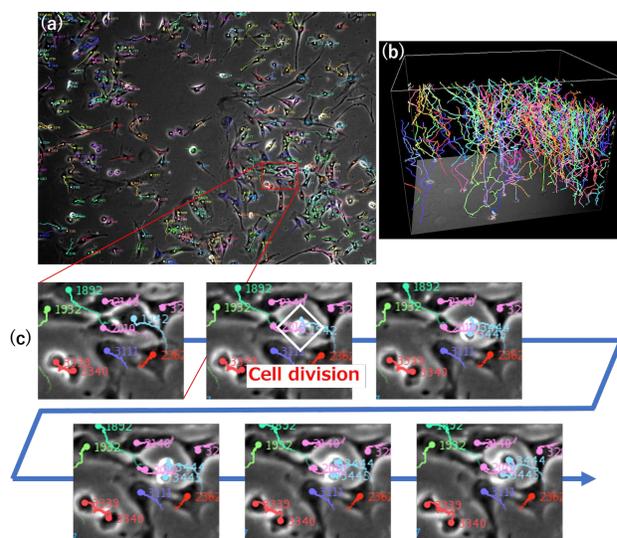


Figure 11. Examples of our tracking results. (a) Entire image. (b) 3D view of estimated cell trajectories. The z-axis indicates the time, and each color indicates the trajectory of a single cell. (c) Enlarged image sequence at the red box in (a), in which our method correctly identified the cell division and tracked all cells.

ence; *i.e.*, if a cell is detected, the association of the cell is always produced, and thus it improved the tracking performance. In the experiments, our method outperformed the state-of-the-art method with over 5% improvement of the tracking performance under various conditions in real biological images. In future work, we will develop an end-to-end tracking method that can estimate the entire cell trajectories from the entire sequence in order to use the global spatial-temporal information on CNN.

## Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP18H05104 and JP18H04738.

## References

- [1] Saad Ullah Akram, Juho Kannala, Lauri Eklund, and Janne Heikkilä. Joint cell segmentation and tracking using cell proposals. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 920–924, 2016.
- [2] Sadeghian Amir, Alahi Alexandre, and Savarese Silvio. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 300–311, 2017.
- [3] Assaf Arbelle and Tammy Riklin Raviv. Microscopy cell segmentation via convolutional lstm networks. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1008–1012, 2019.
- [4] Robert Bensch and Olaf Ronneberger. Cell segmentation and tracking in phase contrast images using graph cut with asymmetric boundary costs. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1220–1223, 2015.
- [5] Ryoma Bise, Kang Li, Sungeun Eom, and Takeo Kanade. Reliably tracking partially overlapping neural stem cells in dic microscopy image sequences. In *International Conference on Medical Image Computing and Computer-Assisted Intervention Workshop (MICCAIW)*, pages 67–77, 2009.
- [6] Ryoma Bise, Yoshitaka Maeda, Mee-hae Kim, and Masahiro Kino-oka. Cell tracking under high confluency conditions by candidate cell region detection-based-association approach. In *Proceedings of Biomedical Engineering*, pages 1004–1010, 2011.
- [7] Ryoma Bise and Yoichi Sato. Cell detection from redundant candidate regions under non-overlapping constraints. *IEEE Trans. on Medical Imaging*, 34(7):1417–1427, 2015.
- [8] Ryoma Bise, Zhaozheng Yin, and Takeo Kanade. Reliable cell tracking by global data association. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1004–1010, 2011.
- [9] Doersch Carl and Zisserman Andrew. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [10] Joe Chalfoun, Michael Majurski, Alden Dima, Michael Halter, Kiran Bhadriraju, and Mary Brady. Lineage mapper: A versatile cell and particle tracker. *Scientific reports*, 6:36984, 2016.
- [11] Eigen David and Fergus Rob. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3994–4003, 2015.
- [12] Zhao Dawei, Fu Hao, Xiao Liang, Wu Tao, and Dai Bin. Multi-object tracking with correlation filter for autonomous vehicle. *Sensors*.
- [13] Ker Elmer D.F., Sho Sanami Seunguen, Eom, and Ryoma Bise *et al.* Phase contrast time-lapse microscopy datasets with automated and manual cell tracking annotations. *Scientific Data*, 2018.
- [14] Yu Fengwei, Li Wenbo, Li Quanquan, Liu Yu, Shi Xiaohua, and Yan Junjie. Multiple object tracking with high performance detection and appearance feature. In *European Conference on Computer Vision (ECCV)*, pages 36–42, 2016.
- [15] Jan Funke, Lisa Mais, Andrew Champion, Natalie Dye, and Dagmar Kainmueller. A benchmark for epithelial cell tracking. In *Proceedings of The European Conference on Computer Vision (ECCV) Workshops*, 2018.
- [16] Gkioxari Georgia, Hariharan Bharath, Girshick Ross, and Malik Jitendra. R-cnns for pose estimation and action detection. In *arXiv*, 2014.
- [17] Junya Hayashida and Ryoma Bise. Cell tracking with deep learning for cell detection and motion estimation in low-frame-rate. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 397–405, 2019.
- [18] Kokkinos Iasonas. Ubertnet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [19] Misra Ishan, Shrivastava Abhinav, Gupta Abhinav, and Hebert Martial. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3994–4003, 2016.
- [20] Valmadre Jack, Bertinetto Luca, Henriques Joao, Vedaldi Andrea, and Torr Philip H.S. End-to-end representation learning for correlation filter based tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2805–2813, 2017.
- [21] Takeo Kanade, Zhaozheng Yin, Ryoma Bise, Seungil Huh, Sungeun Eom, Michael F. Sandbothe, and Mei Chen. Cell image analysis: Algorithms, system and applications. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2011.
- [22] Bo Li, Wei Wu, Zheng Zhu, and Junjie Yan. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [23] Kang Li and Takeo Kanade. Nonnegative mixed-norm preconditioning for microscopy image segmentation. In *International Conference Information Processing in Medical Imaging (IPMI)*, pages 362–373, 2009.
- [24] Kang Li, Eric D Miller, Mei Chen, Takeo Kanade, Lee E Weiss, and Phil G Campbell. Cell population tracking and lineage construction with spatiotemporal context. *Medical image analysis*, 12(5):546–566, 2008.
- [25] Min Liu, Ram Kishor Yadav, Amit Roy-Chowdhury, and G Venugopala Reddy. Automated tracking of stem cell lineages of arabidopsis shoot apex using local graph matching. *The Plant Journal*, 62(1):135–147, 2010.
- [26] Wang Lu, Xu Lisheng, Kim Min-Young, Luca Rigazico, and Yang Ming-Hsuan. Online multiple object tracking via flow and convolutional features. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 3630–3634, 2017.
- [27] Filip Lux and Petr Matula. Dic image segmentation of dense cell populations by combining deep learning and watershed.

- In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 236–239, 2019.
- [28] Katya Mkrtychyan, Damanpreet Singh, Min Liu, V Reddy, A Roy-Chowdhury, and M Gopi. Efficient cell segmentation and tracking of developing plant meristem. In *IEEE International Conference on Image Processing (ICIP)*, pages 2165–2168, 2011.
- [29] Kazuya Nishimura, Ryoma Bise, et al. Weakly supervised cell instance segmentation by propagating from detection response. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 649–657. Springer, 2019.
- [30] Kenji Okuma, Ali Taleghani, Nando De Freitas, James J Little, and David G Lowe. A boosted particle filter: Multitarget detection and tracking. In *European conference on computer vision (ECCV)*, pages 28–39, 2004.
- [31] Al-Kofahi Omar, Radke Richard, Goderie Susan, Shen Qin, Temple Sally, and Badrinath Roysam. Automated cell lineage construction: A rapid method to analyze clonal development established with murine neural progenitor cells.
- [32] Christian Payer, Darko Štern, Thomas Neff, Horst Bischof, and Martin Urschler. Instance segmentation and tracking with cosine embeddings and recurrent hourglass networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 3–11, 2018.
- [33] Markus Rempfler, Sanjeev Kumar, Valentin Stierle, Philipp Paulitschke, Bjoern Andres, and Bjoern H Menze. Cell lineage tracing in lens-free microscopy videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 3–11, 2017.
- [34] Markus Rempfler, Valentin Stierle, Konstantin Ditzel, Sanjeev Kumar, Philipp Paulitschke, Bjoern Andres, and Bjoern H Menze. Tracing cell lineages in videos of lens-free microscopy. *Medical image analysis*, 48:147–161, 2018.
- [35] Henschel Roberto, Marcard Timo von, and Rosenhahn Bodo. Simultaneous identification and tracking of multiple people using video and imus. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2019.
- [36] Martin Schiegg, Philipp Hanslovsky, Bernhard X Kausler, Lars Hufnagel, and Fred A Hamprecht. Conservation tracking. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2928–2935, 2013.
- [37] Sun ShiJie, Akhtar Naveed, Song HuanSheng, Mian S.Ajmal, and Shah Mubarak. Deep affinity network for multiple object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [38] Ihor Smal, Wiro Niessen, and Erik Meijering. Bayesian tracking for fluorescence microscopic imaging. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 550–553, 2006.
- [39] Hang Su, Zhaozheng Yin, Seungil Huh, and Takeo Kanade. Cell segmentation in phase contrast microscopy images via semi-supervised classification over optics-related features. *Medical image analysis*, 17(7):746–765, 2013.
- [40] Xiaoxu Wang, Weijun He, Dimitris Metaxas, Robin Mathew, and Eileen White. Cell segmentation and tracking using texture-adaptive snakes. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 101–104, 2007.
- [41] Xiao Wen, Vallet Bruno, Schindler Konrad, and Paparoditis Nicolas. Simultaneous detection and tracking of pedestrian from panoramic laser scanning data. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2016.
- [42] Zheng Wu, Danna Gurari, Joyce Wong, and Margrit Betke. Hierarchical partial matching and segmentation of interacting cells. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 389–396, 2012.
- [43] Fuxing Yang, Michael A Mackey, Fiorenza Ianzini, Greg Gallardo, and Milan Sonka. Cell segmentation, tracking, and mitosis detection using temporal context. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 302–309, 2005.
- [44] Zhaozheng Yin, Takeo Kanade, and Mei Chen. Understanding the phase contrast optics to restore artifact-free microscopy images for segmentation. *Medical image analysis*, 16(5):1047–1062, 2012.
- [45] Zeng Zhiliang, Xianzhi Li, Kin Ying, and Fu Yu Chi-Wing. Deep floor plan rproceedings of theecognition using a multi-task network with room-boundary-guided attention. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9096–9104, 2019.
- [46] Zibin Zhou, Fei Wang, Wenjuan Xi, Huaying Chen, Peng Gao, and Chengkang He. Joint multi-frame detection and segmentation for multi-cell tracking. *International Conference on Image and Graphics (ICIG)*, 2019.