Weakly-Supervised Cell Tracking via Backward-and-Forward Propagation

Kazuya Nishimura[1], Junya Hayashida[1], Chenyang Wang[2], Dai Fei Elmer Ker[2], and Ryoma Bise[1]

[1] Kyushu University, Fukuoka, Japan
{kazuya.nishimura,bise}@human.ait.kyushu-u.ac.jp
[2] The Chinese University of Hong Kong, Hong Kong

Abstract. We propose a weakly-supervised cell tracking method that can train a convolutional neural network (CNN) by using only the annotation of "cell detection" (*i.e.*, the coordinates of cell positions) without association information, in which cell positions can be easily obtained by nuclear staining. First, we train co-detection CNN that detects cells in successive frames by using weak-labels. Our key assumption is that co-detection CNN implicitly learns association in addition to detection. To obtain the association, we propose a backward-and-forward propagation method that analyzes the correspondence of cell positions in the outputs of co-detection CNN. Experiments demonstrated that the proposed method can associate cells by analyzing co-detection CNN. Even though the method uses only weak supervision, the performance of our method was almost the same as the state-of-the-art supervised method. Code is publicly available in https://github.com/naivete5656/WSCTBFP .

Keywords: Cell tracking, weakly-supervised learning, multi-object tracking, cell detection, tracking, weakly-supervised tracking

1 Introduction

Cell behavior analysis plays an important role in biology and medicine. To create quantitative cell-behavior metrics, cells are often captured with time-lapse images by using phase-contrast microscopy, which is a non-invasive imaging technique, and then hundreds of cells over thousands of frames are tracked in populations. However, it is time-consuming to track a large number of cells manually. Thus, automatic cell tracking is required.

Cell tracking in phase-contrast microscopy has several difficulties compared with general object tracking. First, cells have similar appearences and their shapes may be severely deformed. Second, cells often touch each other and have blurry intercellular boundaries. Third, a cell may divide into two cells (cell mitosis); this is very different from general object tracking. These aspects make it difficult to track cells by using only shape similarity and proximity of cells.

To address such difficulties, the positional relationship of nearby cells is important information to identify the association. The recently proposed CNN-

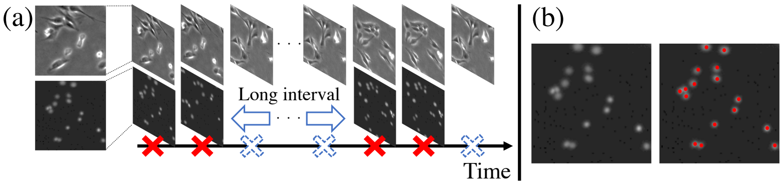

Fig. 1. (a) **Top**: Phase-contrast image sequence. **Bottom**: Fluorescent images with the nucleus stain cell; these were only captured with longer intervals (red crosses) due to phototoxicity. Fluorescent images were not captured on the white crosses. (b) Rough centroid positions (in red) can be easily identified in fluorescent images.

based methods that use such context [31, 13] have outperformed the conventional image-processing-based methods. However, learning-based methods require enough training data including individual cell positions in each frame and their correspondences in successive frames (*i.e.,* cell location and motion). In addition, since the appearance and behaviors of a cell often change the appearance and behaviors of a cell often change may often change depending on condition (*e.g.,* growth-factors, type of microscope), we usually have to prepare a training dataset for each individual case.

On the other hand, there are invasive imaging techniques such as fluorescent imaging to facilitate observation of cells. If we can obtain fluorescent images showing cells whose nuclei are stained (Fig. 1) in addition to the phase-contrast images, the rough centroid positions can be easily detected by using simple image processing techniques. However, because fluorescent imaging damages cells, these images can be only captured for training, not for testing. Moreover, fluorescent images cannot be captured frequently over a long period, since phototoxicity may affect the shapes and migration of the cells. Instead, we can capture fluorescent images only several times in enough long period (Fig. 1) since cells can recover from the damage during the non-invasive imaging period. From such sequences, we can automatically obtain point labels for detection [28]. Although these labels do not include the correspondence information between frames, they can be considered as weak-labels for the tracking task.

In this paper, we propose a weakly supervised cell tracking that can obtain the correspondences from training data for the detection tasks (without association). In order to obtain the association information, we designed a method that has three steps as shown in Fig. 2: (1) Our co-detection CNN is trained to detect cells in successive frames by using the rough cell centroid positions, which are weak-labels for the tracking task but it can be used as supervision for detection. Our key assumption is that co-detection CNN implicitly learns the association. (2) The proposed method performs backward-and-forward propagation to extract associations from co-detection CNN without any ground-truth. When we focus on a particular detection response in the output layer L_{t+1} (*e.g.,* the red region in the left image of Fig. 2 (2-1)), the association problem can be considered to be one of finding the position corresponding with the cell of interest (blue

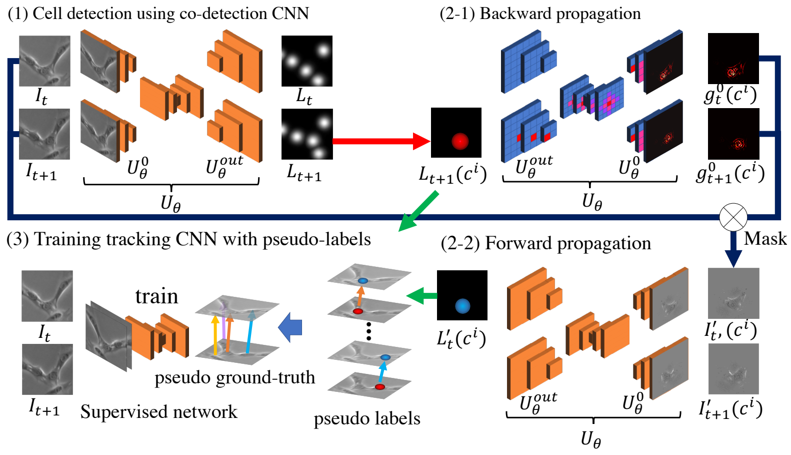

Fig. 2. Overview of our method. (1) co-detection CNN U_θ estimates the position likelihood maps for two successive frames. (2-1) Backward-propagation estimates relevance maps $g_t^0(c^i)$, $g_{t+1}^0(c^i)$ of the cell of interest c^i (red). (2-2) Forward-propagation estimates the cell position likelihood map $L'_t(c^i)$ (blue) with inputting the masked images $I'_t(c^i)$, $I'_{t+1}(c^i)$ which are generated using $g_t^0(c^i)$, $g_{t+1}^0(c^i)$. The pseudo-labels are generated using this estimated regions. (3) The tracking CNN is trained using the pseudo-labels.

region in Fig. 2 (2-2)) from L_t. (3) Using the detection results (1) and association results (2), we can generate the pseudo-training data for the tracking task. We train the cell tracking method [13] with pseudo-training data and a masked loss function that ignores the loss from the false-negative regions, in which we can know the false-negative regions where the cell of interest cannot be associated with any cells in the second step. It is expected that the trained tracking network has better tracking performance compared with the pseudo-labels.

Our main contributions are summarized as follows:

- We propose a weakly-supervised tracking method that can track multiple cells by only using training data for detection. Our method obtains cell association information from co-detection CNN. The association information is used as pseudo-training data for cell tracking CNN.
- We propose a novel network analysis method that can estimate the relation between output and output in two outputs CNN. Our method can extract the positional correspondences of cells from two successive frames by analyzing co-detection CNN.
- We demonstrated the effectiveness of our method using open data and realistic data. In realistic data, we do not use any human annotations. Our method outperformed current methods that do not require training data. In addition, the performance of our method was almost the same as the state-of-the-art supervised method by using only weak supervision.

2 Related work

Cell tracking: Many cell tracking methods have been proposed, which is particle filters [30, 37], active contour [22, 45, 46, 51], and detection-and-association [17, 8, 10, 9, 35, 40, 51]. The detection-and-association methods, which first detect cells in each frame and then solve associations between successive frames, are the most popular tracking paradigm due to the good quality of detection algorithms that use CNNs in the detection step [32, 33, 3, 25]. To associate the detected cells, many methods use hand-crafted association scores based on proximity and shape similarity [17, 51, 10, 35, 40]. To extract the similarity features from images, Payer *et al.* [31] proposed a recurrent hourglass network that not only extracts local features but also memorizes inter-frame information. Hayashida *et al.* [13, 14] proposed a cell motion field that represents the cell association between successive frames and it can be estimated by a CNN. These methods outperform ones that use hand-crafted association scores. However, they require sufficient training data for both detection and association.

Unsupervised or weakly-supervised tracking for general objects: Recently, several unsupervised or weakly-supervised tracking methods have been proposed. To track a single object, correspondence learning have been proposed with several weakly-supervision [49] or unsupervised scenarios [42, 44, 43]. Zhong *et al.* [49] proposed a tracking method that combines the outputs of multiple trackers to improve tracking accuracy in order to address noisy labels (weak labels). The weak label scenario is different from ours. These methods assumed for tracking a single object, and thus these are short to our problem. Several methods have been proposed for multi-object tracking [29, 16]. Nwoye *et al.* [29] proposed a weakly-supervised tracking of surgical tools appearing in endoscopic video. The weak label in this case is a class label of the tool type, and they assumed that one tool appears for each tool type even though several different types of tools appear at a frame. Huang *et al.* [16] tackled a similar problem of semantic object tracking. He *et al.* [15] proposed an unsupervised tracking-by-animation framework. This method uses shape and appearance information for updating the track states of multiple-objects. It assumes that the target objects have different appearances. The above methods may become confused if there are many similar appearance objects in the image; such is the case in cell tracking.

Relevant pixel analysis: Visualization methods have been proposed for analyzing relevant pixels for classification in CNNs [38, 4, 27, 39, 50, 36, 12, 48, 20]. Layer-wise relevance propagation (LRP) [4, 27] and guided backpropagation [39] back-propagate signals from the output layer to the input layer on the basis of the weights and signals in the forward-propagation for inference. Methods based on class activation mapping (CAM) [50, 36, 12], such as Grad-CAM [12], produces the relevance map from CNN using the semantic features right before the fully connected layer for classification. There are several methods that uses such backward operation in a network for instance segmentation [28] and object tracking [23]. For example, Li *et al.* [23] propose a Gradient-Guided Network (GradNet) for a single object tracking that exploits the information in the gradient for template update. Although this method uses the backward operation for guided calculation, the purpose is totally different from ours (analysis of the relevance of the two output layers). These methods assume that they analyze the relationship between the input and output layers but not for two output layers in a multi-branch network.

Unlike the above methods, our method can obtain correspondences between objects in successive frames without training data for the association by analyzing the co-detection CNN, in spite of the challenging conditions wherein many cells having similar appearances migrate.

3 Weakly-supervised cell tracking

3.1 Overview

Fig. 2 shows an overview of the proposed method. The method consists of three parts: 1) cell detection using co-detection CNN that jointly detects cells at successive frames using weak labels (cell position label): it is expected to implicitly learn the association between the frames in addition to detection; 2) backward-and-forward propagation for extracting the association from co-detection CNN: 3) training the cell tracking network using pseudo-labels that were generated using the estimated results in (2); The details of each step are explained as follows.

3.2 Co-detection CNN

In the co-detection task [6], the detection results in frame t can facilitate to detect the corresponding cell in frame $t + 1$ and vice versa. Based on this key observation, we designed co-detection CNN U_θ for jointly detecting cells at the successive frames, in which θ indicates the network parameters. In our problem setup, the rough cell centroid position is obtained from the fluorescent images as training data, but the nuclei position may shift from the ground-truth of the centroid position. Therefore, we follow the cell detection network [28] that mitigates this gap by representing cell positions as a position likelihood map. The position likelihood map can be generated from the rough cell centroid position, where a given cell position becomes an intensity peak and the intensity value gradually decreases away from the peak in accordance with a Gaussian distribution [28]. In contrast to [28] (U-Net [34] architecture), our network has two input-encoders, a common network, and two output-decoders to simultaneously estimates the detection results in successive frames as shown in Fig. 3. The two input-encoders have shared weights, and these extract the cell appearance features from the inputted successive images I_t, I_{t+1}. The features are concatenated and input into a common network that has a U-Net architecture. We consider that the common network performs co-detection and it implicitly learns the cell association. Finally, the output-decoders decode the extracted co-detection features into the cell position likelihood maps; the layers of the input and output

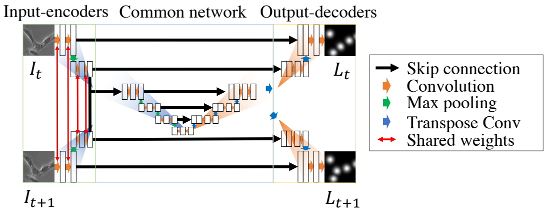

Fig. 3. Architecture of co-detection CNN.

networks have skip connections to adjust the local positions similar to U-Net. The loss function $Loss_{CD}$ for co-detection CNN is the sum of the Mean Square Errors (MSE) of the likelihood maps of the two frames:

$$Loss_{CD} = MSE(L_t - \hat{L}_t) + MSE(L_{t+1} - \hat{L}_{t+1}), \tag{1}$$

where $\hat{L}_t$, $\hat{L}_{t+1}$ are the ground-truths of the cell position likelihood map of each frame and L_t, L_{t+1} are the estimated maps. In the inference, the peaks in the estimated map are the detected cell positions.

3.3 Backward-and-Forward propagation

Next, in accordance with our assumption that co-detection CNN implicitly learns the association, we extract the cell association information from co-detection CNN U_θ. Here, we will focus on a particular detection response in the output layer L_{t+1} (*e.g.,* the red regions in Fig. 2 (2-1)). The association problem can be considered to be one of finding the position (blue region) corresponding to the cell of interest from L_t. We propose the following backward-and-forward propagation for this task.

Backward propagation: Fig. 2 (2-1) illustrates the backward-propagation process on U_θ for the cell of interest c^i that is selected from frame $t+1$. In this step, we extract the relevance maps that are expected to relevant to producing the detection response of interest by using guided backpropagation (GB) [39]. For this process, we modified the weakly-supervised instance segmentation method proposed by Nishimura [28], which extracts individual relevant cell regions of a particular cell in U-Net for a single image. Different from [28], in our case, two relevance maps $g^0_t(c^i)$, $g^0_{t+1}(c^i)$ are extracted from a single output $L_{t+1}(c^i)$.

The GB back propagates the signals from the output layer U^{out}_θ to the input layer U^0_θ by using the trained parameters (weights) θ in the network. In our method, to obtain the individual relevant cell regions of a particular cell, we first initialize the cell position likelihood map $L_{t+1}(c^i)$ for each cell of interest c^i, in which all regions outside the cell region substitute 0 (Fig. 2 (2-1)). The region within radius r from the coordinate of the cell of interest c^i is defined as $S(c^i)$. Then, the relevant pixels of each cell of interest were obtained by backpropagation from $S(c^i)$. The backpropagating signals are propagated to both layers at the branch of the input-encoders. The red nodes of the intermediate layers in the network in Fig. 2 (2-1) show the illustration of the back-propagation process. This backward process is performed for each cell $i = 1, ..., N$, where N is the number of cells.

It is expected that the corresponding cell regions have positive values in the relevance maps. However, regions of the outside the cell of interest may also have values in the relevance map. This adversely affects the process of extracting the cell association. Therefore, we compare the pixel values of $g^0_{p,t}(c^i)$ $(i = \{1, ..., N\})$ for all cells, where pixel p corresponds to c^i if it takes the maximum value among the cells The maximum projection of the p-th pixel for c^i at frame t can be formalized as:

$$g'_{p,t}(c^i) = \psi_p(t, i, \arg\max_k g^0_{p,t}(c^k)), \tag{2}$$

$$\psi_p(t, i, k) = \begin{cases} g^0_{p,t}(c^i) & if\ (k = i), \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where, p is the p-th pixel on the relevance map. $g'_{t+1}(c^i)$ is calculated by same manner. By applying maximum projection to all cells, we get the maximum projection relevance maps $g'_t(c^i)$ and $g'_{t+1}(c^i)$ for each cell.

Forward propagation: This step estimates the corresponding cell position likelihood map $L'_t(c^i)$ by using the relevance maps $g'_t(c^i)$ and $g'_{t+1}(c^i)$ (Fig. 2 (2-2)). The high value pixels in $g'_t(c^i)$, $g'_{t+1}(c^i)$ show the pixels that contribute to detect c^i. It indicates that co-detection CNN U_θ is able to detect the corresponding cell position from only these relevant pixels in the input images.

We generate masked images $I'_t(c^i)$, $I'_{t+1}(c^i)$ that only have values at the high relevance pixels of only the cell of interests c^i. In order to generate it, we first initialize the images so that it has the background intensities of the input image. Then, we set the pixel values in the initialized image as the input image intensity if the value of the p-th pixel of $g'_t(c^i)$ is larger than a threshold th.

$$I'_{p,t}(c^i) = \begin{cases} I_{p,t}(c^i) & \text{if } g'_{p,t}(c^i) > th, \\ B_{p,t}(c^i) & \text{otherwise,} \end{cases} \tag{4}$$

where the background image at frame t B_t is estimated by using quadratic curve fitting [47] and p is the p-th pixel on I'. $I'_{t+1}(c^i)$ is also made by the same manner. The corresponding cell position likelihood map $L'_t(c^i) = U_\theta(I'_t(c^i), I'_{t+1}(c^i))$ can be obtained by inputting the masked images $I'_t(c^i)$, $I'_{t+1}(c^i)$ to U_θ.

The estimated map $L'_t(c^i)$ indicates the i-th detection response at t that corresponds to the detected cell at frame t. It is expected that the high intensity region in $L'_t(c^i)$ is expected to appears on the same region of either cell detection result in L_t. To obtain the cell position at t corresponding to the i-th cell position at $t + 1$, we perform one-by-one matching by using linear programming between these two maps, in which we use a simple MSE of intensities for the matching score. This one-by-one matching may correspond either of the cell position even if the estimated response signal is too small. We omit the low confidence associations in order to keep the precision of the pseudo-labels high enough. If the

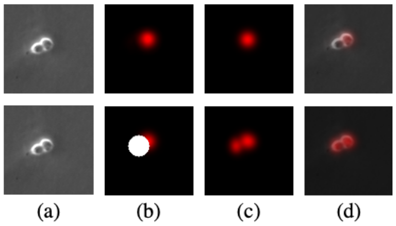

Fig. 4. Top: Example without using the masked loss. **Bottom:** Example with using the masked loss. (a) phase-contrast image, (b) pseudo-training data, (c) output of trained network, and (d) overlapping images. The white region in (b) indicates the region that ignores the loss.

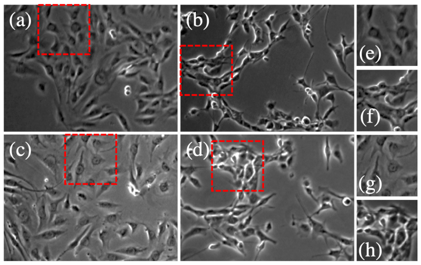

Fig. 5. Example images on four calture condtions. (a) Control, (b) FGF2, (c) BMP2, (d) FGF2+BMP2, (e)-(h) are enlarged images of the red box in (a)-(d).

matching score is less than the threshold th_{conf}, we define it as the low confidence association. By omitting low confidence associations, the result includes some false-negative. One of the interesting points is that we can know where the low confidence region at $t+1$ is since we explicitly give the region of the cell of interest $S_{t+1}(c^i)$. If the cell c^i is not associated with any cell, we add the pixels of the cell $S_{t+1}(c^i)$ to the set of unassociated cell region $\mathbf{\Gamma}$. Finally, we obtain the cell position and association in successive frames and there will use as pseudo-labels in the next step.

3.4 Training tracking CNN using pseudo-labels

It is known that the generalization performance and estimation speed can be improved by training a CNN using pseudo-labels in weakly-supervised segmentation tasks [5, 19, 24, 1, 2]; we took this approach for our tracking task. We train a state-of-the-art cell tracking network called MPM-Net [14] with the pseudo-labels. The MPM-Net estimates Motion and Position Map (MPM) that simultaneously represents the cell positions and their association between frames from inputting the images at successive frames. The advantage of this method is that it can extract the features of the spatio-temporal context about nearby cells from the inputted entire images for association and detection. We generate the pseudo-training data for MPM using obtained cell position and association in Sec. 3.3.

In the previous step, we obtained the high confidence pseudo-labels and a set of unassociated regions $\mathbf{\Gamma}$. The top row in Fig. 4(b) shows an example of pseudo-training data directly generated using only the high confidence labels. In this example, a cell divides two cells. However, the mother cell was associated with one of the child cells and the other was not (false negative) due to one-by-one matching. If we train the network using such noisy labels, this non-associated cell region affects the learning. Indeed, the non-associated region was not detected due to over-fitting. Fig. 4(c) shows the output of the trained network that only detects one cell by over-fitting. To avoid this problem, we train the network with the masked loss function that ignores the loss from the false-negative regions where the cell did not correspond to any detection responses due to its low confidence. The masked loss is formulated as:

$$Loss_{mask} = \begin{cases} 0 & \text{if } \boldsymbol{p} \in \boldsymbol{\Gamma}, \\ Loss_{ori} & \text{otherwise}, \end{cases} \tag{5}$$

where $\boldsymbol{\Gamma}$ is the set of the ignoring regions that contain the unassociated cell regions, $\boldsymbol{p}$ is a pixel in $\boldsymbol{\Gamma}$, $Loss_{ori}$ is the original loss for MPM-Net [14]. As shown in the white circle in bottom row of Fig. 4(b), the false-negative region is not calculated in the masked loss. This effectively avoid over-fitting and correctly estimate the cell position likelihood map as shown in the bottom row of Fig. 4(c).

4 Experiment

4.1 Data set and experimental setup

We evaluated our method on an open data set [18] that contains time-lapse sequences captured phase-contrast microscopy[3]. In the data set, the mybolast cells were cultured under four growth factor conditions: (a) Control, (b) FGF2, (c) BMP2, and (d) FGF2+BMP2 (Fig. 5). Each sequence consists of 780 frames, with a 5 minute interval between consecutive frames. The resolution of each image is 1392× 1040 pixels. There are four sequences for each condition and the total number of sequences is 16. The rough cell centroid positions are annotated with the cell ID. In one of the BMP2 sequence, all cells are annotated. For the other sequences, three cells were randomly selected at the beginning of the sequence and then their descendants were annotated. The total number of annotated cells in the 16 sequences is 135859. We used one of the BMP2 sequence as the training data for co-detection CNN and the other sequences were used as the test data. In the training process, we only used the cell position coordinates as weak labels. The task was challenging because the training was only weakly-supervised and the appearances of the cells in the test data differed from those in the training data (see Fig. 5). To train co-detection CNN and MPM, we used Adam [21] optimizer with learning rate 10^{-3}. We set the threshold th in Eq. 4 to 0.01, the low confidence association threshold th_{conf} to 0.5, and $r = 18$ in all experiments; these parameters were decided using validation data and were not sensitive.

4.2 Performance of cell tracking on open data set

We compared our method with five other methods by using the open data set [18]. Since our method only requires weak-supervision, we selected three

[3] The data [18] is more challenging as a tracking task compared with ISBI Cell Tracking Challenging [41, 26] that more focused on segmentation task, since the cells often partially overlapped and the boundary of cells is ambiguous.

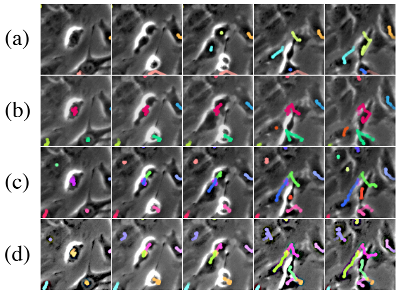

Fig. 6. Examples of tracking results of (a) GDA, (b) CMF, (c) MPM, and (d) ours. The horizontal axis indicates the time.

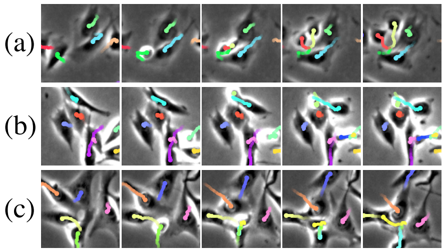

Fig. 7. Examples of tracking results under each conditions: (a) Control, (b)FGF2, and (c) FGF2+BMP2.

methods that do not use association information; 1) asymmetric graphcut (A-Graph) [7] that segments cell regions using asymmetric graph-cut: it was trained with the small amount of additional ground-truth for segmentation; 2) Fogbank [11] that segments cell regions using image processing: the hyper-parameters were tuned using the validation data; 3) global data association (GDA) [10] that segments cell regions by physical-model-based method [47] and then performs spatial-temporal global data association: the hyper-parameters were tuned using validation data. In addition, in order to show that the performance of our method is comparable with the SOTA (state-of-the-art), we evaluated two supervised tracking methods that require the ground-truth of the cell position and association; 4) cell motion field (CMF) [13] that estimates the cell motion and position separately; 5) motion and position map (MPM) [14] that estimates the motion and position map, which achieved the SOTA performance. In addition, to confirm the effectiveness of the masked loss, we also compared with our method without the masked loss (Ours w/o ml).

We used the association accuracy and target effectiveness as following the paper that proposed the MPM [14]. Each target was first assigned to a track (estimation) for each frame. The association accuracy indicates the number of true positive associations divided by the number of true positive associations in the ground-truth. If cell A switches into B, and B into A, we count two false-positive (A→B, B→A) and two false-negatives (no A→A, B→B). The target effectiveness was computed as the number of the assigned track observations over the total number of frames of the target after assigning each target to a track that contains the most observations from that ground-truth. It indicates how many frames of targets are followed by computer-generated tracks. This metric is a stricter than the association accuracy. If a switching error occurs in the middle of the trajectory, the target effectiveness is 0.5.

Table 1 shows the results of the performance comparison. Our method outperformed the other weakly or unsupervised methods (A-Graph [7], Fogbank [11], GDA [10]) and achieved comparable results with state-of-the-art supervised methods (MPM [14]). Even though our method used only weak-supervision, it

Table 1. Tracking performance in terms of association accuracy (AA) and target effectiveness (TE) on open data set [18]. Su. indicates the condition of the training data: weak-supervision (W), un-supervision (U), and fully-supervision (F). The best and second best are denoted by boldface and the best one is underlined. '*' indicates the culture condition is the same as in the training data.

Method	Su.	*BMP2		FGF2		Control		FGF2+ BMP2		Ave.	
		AA	TE	AA	TE	AA	TE	AA	TE	AA	TE
A-Graph [7]	W	0.801	0.621	0.604	0.543	0.499	0.448	0.689	0.465	0.648	0.519
Fogbank [11]	U	0.769	0.691	0.762	0.683	0.650	0.604	0.833	0.587	0.753	0.641
GDA [10]	U	0.855	0.788	0.826	0.733	0.775	0.710	0.942	0.633	0.843	0.771
Ours w/o ml	W	0.979	0.960	0.950	0.861	0.917	0.786	0.972	0.880	0.954	0.873
Ours	W	**0.982**	**$\underline{0.970}$**	**$\underline{0.955}$**	**$\underline{0.869}$**	**0.926**	**0.806**	**0.976**	**$\underline{0.911}$**	**0.960**	**$\underline{0.881}$**
CMF [13]	F	0.958	0.939	0.866	0.756	0.884	0.761	0.941	0.841	0.912	0.822
MPM [14]	F	**$\underline{0.991}$**	**0.958**	**0.947**	**0.803**	**$\underline{0.952}$**	**$\underline{0.829}$**	**$\underline{0.987}$**	**$\underline{0.911}$**	**$\underline{0.969}$**	**0.875**

Table 2. Detection performance.

Method	*BMP2 sparse	*BMP2 medium	*BMP2 dense	Control	FGF2	FGF2 +BMP2
Nishimura [28]	0.998	0.978	0.977	0.922	0.924	**0.962**
Ours (5 min. int.)	**0.999**	0.983	**0.980**	0.923	**0.928**	0.945
Ours (25 min. int.)	0.998	**0.984**	0.978	**0.926**	0.911	0.946

outperformed that of the supervised MPM in FGF2. We consider that the MPM may be over-fitted to the condition of the training data (BMP2), and thus its performance may decrease since the cell appearance in FGF2 is different from that in BMP2 as shown in Fig. 5. In addition, the results show that the masked loss slightly improved the performance compared with 'Ours w/o ml'. Fig. 6 (c) shows examples of tracking results under BMP2[4]. Although GDA did not detect the brighter cell and CMF did not identify the newly born cells after cell mitosis, our method successfully tracked almost all the cells as the same with MPM.

4.3 Ablation study

Next, we performed an ablation study to evaluate the performance of the co-detection CNN, backward-and-forward propagation, and re-training individually. In this ablation study, in order to confirm the robustness in various conditions, we additionally added annotations for other three conditions (Control, FGF2, FGF2+BMP2) since only some of the cells were annotated under these three conditions in the original data. Then, we evaluated each step of our method under these three conditions.

Co-detection CNN: We first evaluated our co-detection CNN against the method proposed by Nishimura *et al.* [28] that estimates the cell position likeli-

[4] Since the tracking results of A-Graph and Fogbank were very worse, we omitted their results on these figures due to the page limitation.

Table 3. Association performance of backward-and-forward propagation.

Interval	Metrics	*BMP2 sparse	*BMP2 medium	*BMP2 dense	Control	FGF2	FGF2 +BMP2
5 min.	Precision	0.999	0.989	0.992	0.971	0.966	0.964
	Recall	0.997	0.976	0.957	0.849	0.844	0.900
	F1-score	0.998	0.982	0.974	0.906	0.901	0.931
25 min.	Precision	0.998	0.982	0.974	0.906	0.901	0.931
	Recall	0.961	0.975	0.960	0.849	0.753	0.902
	F1-score	0.979	0.981	0.976	0.904	0.830	0.930

Table 4. Comparison of backward-and-forward propagation (BF) with MPM trained by backward-and-forward propagation (T). AA: association accuracy, TE: target effectiveness.

Met.	*BMP2 sparse		*BMP2 medium		*BMP2 dense		Control		FGF2		FGF2 +BMP2	
	AA	TE	AA	TE	AA	TE	AA	TE	AA	TE	AA	TE
BF	0.983	0.968	0.973	0.962	0.840	0.914	0.826	0.765	**0.794**	**0.672**	0.955	0.945
T	**0.993**	**0.976**	**0.980**	**0.974**	**0.970**	**0.969**	**0.858**	**0.800**	0.773	0.640	**0.982**	**0.970**

hood map of a single image. In addition, in order to demonstrate the robustness for cell migration speed since the speed is depending on the cell types and time-interval, we also evaluated two intervals (5 and 25 minutes), in which the speed in 25 min is much faster than that in 5 min. We used F1-score as the detection performance metric. Table 2 shows the results. Our co-detection CNN performed almost as well as the state-of-the-art method under all conditions (BMP2-sparse, BMP2-medium, BMP2-dense, Control, FGF2, FGF2+FGF2). In addition, the results show that our method was robust to the different cell migration speeds, since the performances of both interval conditions are almost the same. Fig. 8 shows examples in which co-detection CNN improved the detection results. In the upper case, the cell shape is ambiguous at t but it is more clear at $t + 1$, co-detection CNN uses these two image and it may facilitate to detect the cell. In the bottom case, a tips of cell (noise) appears at both frames. Since the noise traveled a large distance, co-detection CNN may reduce over-detections.

Backward-and-forward propagation: Next, we evaluated the association performance of Backward-and-Forward propagation (BF-prop). We used precision, recall, F1-score of association accuracy as the performance metrics. Fig. 9 shows examples. L indicates the output of co-detection CNN given two input images I at t and $t+1$. I' indicates the masked image generated using the relevance map produced by backward propagation. L' indicates the estimated likelihood map by forward propagation by inputting I'. In both cases, the backward propagation could obtain the target cell regions and forward propagation successfully estimated the detection map of the corresponding cell. Under all conditions, backward-and-forward propagation performed association accurately (over 90% in the terms of F1-score as shown in Table 3). As discussed in Sec. 3.3, precision

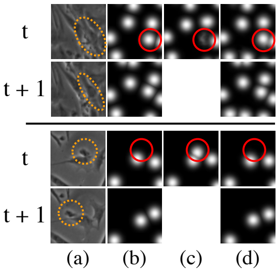

Fig. 8. Example results of detection. (a) phase-contrast, (b) ground-truth, (c) Nishimura [28], and (d) Ours.

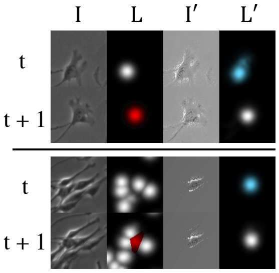

Fig. 9. Example results of BF-prop. The red regions are the cell region of interest, and the blue regions are the estimated corresponding cell.

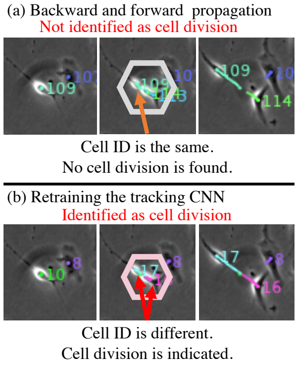
(a) Backward and forward propagation
Not identified as cell division
109
109
109
114
Cell ID is the same.
No cell division is found.
(b) Retraining the tracking CNN
Identified as cell division
10
8
17
16
Cell ID is different.
Cell division is indicated.

Fig. 10. Example results from (a) BF-prop, (b) retrained MPM-net.

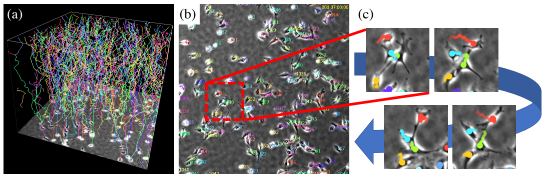

Fig. 11. Examples of tracking result on realistic dat. (a) 3D view of estimated cell trajectories. The z-axis is the time, and each color indicates the trajectory of a cell. (a) Entire image. (c) Sequence of enlarged images at the red box in (b).

Table 5. Quantitative evaluation. AA: association accuracy, TE: target effectiveness

Method	AA	TE
A-Graph [7]	0.216	0.169
Fogbank [11]	0.695	0.321
GDA [10]	0.773	0.527
Ours	**0.857**	**0.804**

is more important than recall when using the pseudo-labels. BF-prop achieved higher precision than recall on all data-sets. In addition, we conducted evaluations using the different intervals (5 and 25 min.). The results for 5 min. were slightly better than those in 25 min, but not significantly.

Training tracking CNN: To show the effectiveness of retraining the tracking CNN, we compared the trained CNN with the results of BF-prop in terms of the same tracking metrics (AA and TE). As shown in Table 4, the retraining improved the performance under almost conditions except FGF2. The important thing is that although the BF-prop only tracks one of the cell when a cell divided two cells, the retraining could identify a cell division since the masked loss helped to detect the another cell of the divided two cells. Fig. 10 shows the example of the cell division case. The two new cells were successfully identified by MPM-Net and new IDs were assigned to them. In contrast, BF-prop tracked only one of them continuously and did not identify cell division.

4.4 Cell tracking without any human annotation

In this section, we consider a more realistic scenario when pairs of phase-contrast and fluorescent microscopy images for training and the test image sequence that was captured by only phase-contrast microscopy were provided by biologists without any human annotation. In order to confirm that our method can perform such a realistic scenario, we also prepared a data-set with this problem setup. In this experiment, 86 pairs of phase-contrast and fluorescent images were given as the training data, and the 95 images was given as the test data, in which the cell appearance is different from the open data set we used in the previous section.

In this setting, our method could perform tracking in four steps. (1) We trained the detection CNN with phase-contrast images using the ground-truth of detection automatically generated from the given fluorescent image(the procedure was the same as that of Nishimura [28].). Then, we generated co-detection pseudo labels. (2) We trained co-detection CNN with the generated detection pseudo labels, and (3) generated the pseudo-labels for association by BF-prop. (4) We trained the MPM-net using the pseudo-labels and applied it to the test data. In the evaluation, we also compared our method with un-supervised and weakly-supervised tracking methods; A-Graph [7], Fogbank [11], GDA [10]. Here, we could not compare with the supervised method on this scenario since there was no supervised training data.

Table 5 shows the tracking results in terms of association accuracy (AA) and target effectiveness (TE). Our method outperformed the other methods on both metrics. Our method achieved an 8% improvement in association accuracy and 28% improvement in target effectiveness compared with the second best. As shown in Fig. 11, our method can track many cells without supervised annotation. These results show that our method can effectively use weak labels and obtain good tracking results.

5 Conclusion

We proposed a weakly-supervised tracking method that can track multiple cells by using only training data for detection. The method first trains co-detection CNN that detects cells in successive frames by using weak supervision. Then, the method estimates association from co-detection CNN by using our novel backward-and-forward propagation method on the basis of the key assumption that co-detection CNN implicitly learns the association. The association is used as pseudo-labels for a state-of-the-art tracking network (MPM-net). Our method outperformed the compared methods and achieved comparable results to those of supervised state-of-the-art methods. In addition, we demonstrated the effectiveness of our method in a realistic scenario in which the tracking network was trained without any human annotations.

Acknowledgement: This work was supported by JSPS KAKENHI Grant Number 20H04211.

References

1. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: CVPR. pp. 2209–2218 (2019)
2. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: CVPR. pp. 4981–4990 (2018)
3. Akram, S.U., Kannala, J., Eklund, L., Heikkilä, J.: Joint cell segmentation and tracking using cell proposals. In: ISBI. pp. 920–924 (2016)
4. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130140 (2015)
5. Bansal, A., Chen, X., Russell, B., Gupta, A., Ramanan, D.: Pixelnet: Representation of the pixels, by the pixels, and for the pixels. arXiv:1702.06506 (2017)
6. Bao, S.Y., Xiang, Y., Savarese, S.: Object co-detection. In: ECCV. pp. 86–101 (2012)
7. Bensch, R., Olaf, R.: Cell segmentation and tracking in phase contrast images using graph cut with asymmetric boundary costs. In: ISBI. pp. 1220–1223 (2015)
8. Bise, R., Li, K., Eom, S., Kanade, T.: 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). pp. 67–77 (2009)
9. Bise, R., Maeda, Y., Kim, M.h., Kino-oka, M.: Cell tracking under high confluency conditions by candidate cell region detection-based-association approach. In: Biomedical Engineering. pp. 1004–1010 (2013)
10. Bise, R., Yin, Z., Kanade, T.: Reliable cell tracking by global data association. In: ISBI. pp. 1004–1010 (2011)
11. Chalfoun, J., Majurski, M., Dima, A., Halter, M., Bhadriraju, K., Brady, M.: Lineage mapper: A versatile cell and particle tracker. Scientific reports **6**, 36984 (2016)
12. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, Vineeth N, B.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: WACV. pp. 839–847 (2018)
13. Hayashida, J., Bise, R.: Cell tracking with deep learning for cell detection and motion estimation in low-frame-rate. In: MICCAI. pp. 397–405 (2019)
14. Hayashida, J., Nishimura, K., Bise, R.: Mpm: Joint representation of motion and position map for cell tracking. In: CVPR, pp. 3823-3832 (2020)
15. He, Z., Li, J., Liu, D., He, H., Barber, D.: Tracking by animation: Unsupervised learning of multi-object attentive trackers. In: CVPR. pp. 1318–1327 (2019)
16. Huang, K., Shi, Y., Zhao, F., Zhang, Z., Tu, S.: Multiple instance deep learning for weakly-supervised visual object tracking. Signal Processing: Image Communication p. 115807 (2020)
17. Kanade, T., Yin, Z., Bise, R., Huh, S., Eom, S., Sandbothe, M.F., Chen, M.: Cell image analysis: Algorithms, system and applications. In: WACV. pp. 374–381 (2011)
18. Ker, E., Eom, S., Sanami, S., Bise, R., Pascale, C., Yin, Z., Huh, S., Osuna-Highley, E., N. Junkers, S., J. Helfrich, C., Yongwen Liang, P., Pan, J., Jeong, S., S. Kang, S., Liu, J., Nicholson, R., F. Sandbothe, M., Van, P., Liu, A., Campbell, P.: Phase contrast time-lapse microscopy datasets with automated and manual cell tracking annotations. Scientific Data **5**, 180237 (11 2018). https://doi.org/10.1038/sdata.2018.237

19. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: CVPR. pp. 876–885 (2017)
20. Kindermans, P.J., Schütt, K.T., Alber, M., Müller, K.R., Erhan, D., Kim, B., Dähne, S.: Learning how to explain neural networks: Patternnet and patternattribution. In: International Conference on Learning Representations (2018)
21. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
22. Li, K., Miller, E.D., Chen, M., Kanade, T., Weiss, L.E., Campbell, P.G.: Cell population tracking and lineage construction with spatiotemporal context. Medical image analysis **12**(5), 546–566 (2008)
23. Li, P., Chen, B., Ouyang, W., Wang, D., Yang, X., Lu, H.: Gradnet: Gradient-guided network for visual object tracking. In: ICCV. pp. 6162–6171 (2019)
24. Li, Q., Arnab, A., Torr, P.H.: Weakly-and semi-supervised panoptic segmentation. In: ECCV. pp. 102–118 (2018)
25. Lux, F., Matula, P.: Dic image segmentation of dense cell populations by combining deep learning and watershed. In: ISBI. pp. 236–239 (2019)
26. Maška, M., Ulman, V., Svoboda, D., Matula, P., Matula, P., Ederra, C., Urbiola, A., España, T., Venkatesan, S., Balak, D.M., et al.: A benchmark for comparison of cell tracking algorithms. Bioinformatics **30**(11), 1609–1617 (2014)
27. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition **65**, 211–222 (2017)
28. Nishimura, K., Bise, R., et al.: Weakly supervised cell instance segmentation by propagating from detection response. In: MICCAI. pp. 649–657 (2019)
29. Nwoye, C.I., Mutter, D., Marescaux, J., Padoy, N.: Weakly supervised convolutional lstm approach for tool tracking in laparoscopic videos. International journal of computer assisted radiology and surgery **14**(6), 1059–1067 (2019)
30. Okuma, K., Taleghani, A., De Freitas, N., Little, J.J., Lowe, D.G.: A boosted particle filter: Multitarget detection and tracking. In: ECCV. pp. 28–39 (2004)
31. Payer, C., Štern, D., Neff, T., Bischof, H., Urschler, M.: Instance segmentation and tracking with cosine embeddings and recurrent hourglass networks. In: MICCAI. pp. 3–11 (2018)
32. Rempfler, M., Kumar, S., Stierle, V., Paulitschke, P., Andres, B., Menze, B.H.: Cell lineage tracing in lens-free microscopy videos. In: MICCAI. pp. 3–11 (2017)
33. Rempfler, M., Stierle, V., Ditzel, K., Kumar, S., Paulitschke, P., Andres, B., Menze, B.H.: Tracing cell lineages in videos of lens-free microscopy. Medical image analysis **48**, 147–161 (2018)
34. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
35. Schiegg, M., Hanslovsky, P., Kausler, B.X., Hufnagel, L., Hamprecht, F.A.: Conservation tracking. In: ICCV. pp. 2928–2935 (2013)
36. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV. pp. 618–626 (2017)
37. Smal, I., Niessen, W., Meijering, E.: Bayesian tracking for fluorescence microscopic imaging. In: ISBI. pp. 550–553 (2006)
38. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv:1706.03825 (2017)
39. Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: ICLRW (2015)

40. Su, H., Yin, Z., Huh, S., Kanade, T.: Cell segmentation in phase contrast microscopy images via semi-supervised classification over optics-related features. Medical image analysis **17**(7), 746–765 (2013)
41. Ulman, V., Maška, M., Magnusson, K.E., Ronneberger, O., Haubold, C., Harder, N., Matula, P., Matula, P., Svoboda, D., Radojevic, M., et al.: An objective comparison of cell-tracking algorithms. Nature methods **14**(12), 1141 (2017)
42. Vondrick, C., Shrivastava, A., Fathi, A., Guadarrama, S., Murphy, K.: Tracking emerges by colorizing videos. In: ECCV. pp. 391–408 (2018)
43. Wang, N., Song, Y., Ma, C., Zhou, W., Liu, W., Li, H.: Unsupervised deep tracking. In: CVPR. pp. 1308–1317 (2019)
44. Wang, X., Jabri, A., Efros, A.A.: Learning correspondence from the cycle-consistency of time. In: CVPR. pp. 2566–2576 (2019)
45. Wang, X., He, W., Metaxas, D., Mathew, R., White, E.: Cell segmentation and tracking using texture-adaptive snakes. In: ISBI. pp. 101–104 (2007)
46. Yang, F., Mackey, M.A., Ianzini, F., Gallardo, G., Sonka, M.: Cell segmentation, tracking, and mitosis detection using temporal context. In: MICCAI. pp. 302–309 (2005)
47. Yin, Z., Kanade, T., Chen, M.: Understanding the phase contrast optics to restore artifact-free microscopy images for segmentation. Medical image analysis **16**(5), 1047–1062 (2012)
48. Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. International Journal of Computer Vision **126**(10), 1084–1102 (2018)
49. Zhong, B., Yao, H., Chen, S., Ji, R., Chin, T.J., Wang, H.: Visual tracking via weakly supervised learning from multiple imperfect oracles. Pattern Recognition **47**(3), 1395–1410 (2014)
50. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929 (2016)
51. Zhou, Z., Wang, F., Xi, W., Chen, H., Gao, P., He, C.: Joint multi-frame detection and segmentation for multi-cell tracking. In: ICIG. pp. 435–446 (2019)