Joint Multi-frame Detection and Segmentation for Multi-cell Tracking

Zibin Zhou, Fei Wang, Wenjuan Xi, Huaying Chen
Peng Gao, and Chengkang He

Harbin Institute of Technology, Shenzhen

Abstract. Tracking living cells in video sequence is difficult, because of cell morphology and high similarities between cells. Tracking-by-detection methods are widely used in multi-cell tracking. We perform multi-cell tracking based on the cell centroid detection, and the performance of the detector has high impact on tracking performance. In this paper, UNet is utilized to extract inter-frame and intra-frame spatio-temporal information of cells. Detection performance of cells in mitotic phase is improved by multi-frame input. Good detection results facilitate multi-cell tracking. A mitosis detection algorithm is proposed to detect cell mitosis and the cell lineage is built up. Another UNet is utilized to acquire primary segmentation. Jointly using detection and primary segmentation, cells can be fine segmented in highly dense cell population. Experiments are conducted to evaluate the effectiveness of our method, and results show its state-of-the-art performance.

Keywords: Multi-frame · Segmentation · Joint · Multi-object · Cell tracking.

1   Introduction

Multi-cell tracking in image sequence is valuable for stem cell research, tissue engineering, drug discovery and proteomics [1]. Researchers can construct cell lineage trees and analyze cell morphology based on cell tracking results [2].

Cell tracking is more challenging than general object tracking. Firstly, cells may be deformed, such as elongation, contraction, and swelling [3]. Secondly, there is a very high similarity between cells. Cells of same kind have the same internal structure, and they are difficult to be distinguished through their appearance. In addition, the irregular motion of cells, the mitotic behavior, the complexity of the background, and the interference of other impurities also increase the challenge.

The development of deep learning in recent years has greatly promoted the progress in computer vision. For example, the performance of ResNet [4] exceeded the performance of humans on the ImageNet test set. Cell tracking has evolved from contour evolution, filtering templates, to tracking-by-detection methods [2]. Researchers continue to improve the robustness of algorithms.

There is no such a multi-cell tracking algorithm that can perform well in all varieties of video sequences. For example, Ref. [5] can perform segmentation and tracking very well when cells are large, but not well when cells are small. Highly dense small cells are apt to be missed during tracking, and lead to tracking errors. To solve this challenge, we propose an algorithm that jointly using detection and segmentation for multi-cell tracking. Our method is composed of four portions: cell centroid detection with multi-frame images, primary multi-cell tracker, primary cell segmentation, fine segmentation. Each portion will be detailed in Sect. 3. The contributions of our work can be summarized as follows:

- Multi-frame as input to UNet [6] is proposed, which helps the network to extract spatio-temporal information. Detection performance of mitotic cells is improved, therefore detection performance of mitosis is improved by the mitosis detection algorithm when tracking.
- A fine cell segmentation algorithm is proposed for tracking highly dense small cells. By jointly using primary tracking results of cell centroid detection and primary cell segmentation results, we achieve a new state-of-the-art performance on dataset Fluo-Hela [7].

Effectiveness of our method is evaluated with Cell Tracking Benchmark [7] of Cell Tracking Challenge (www.celltrackingchallenge.net). Performance metrics include tracking accuracy, segmentation accuracy, and a combination performance of both.

2   Related works

Tracking-by-detection methods are widely used in multi-object tracking [8, 9], as well as in multi-cell tracking. Starting with detecting or segmenting cells in a video sequence, these methods establish temporal associations for cells in frames to frames. Detection performance has high impact on tracking performance [10]. As long as good detection results are available, the tracking problem can be simplified [8]. This paper proposes a tracking-by-detection method, which focuses on detection and segmentation. Detection or segmentation in tracking-by-detection methods will be briefly reviewed as below.

Ciresan et al. [11] proposed to use neural networks for segmentation of microbial images in the early days. They use a neural network as a pixel classifier to segment the biological neuron membrane. The network inference must be run on patch-by-patch separately. Unfortunately patches overlap each other, a large amount of computation redundancy occurs, therefore calculation is quite slow.

Ronneberger et al. [6] proposed the semantic segmentation network, i.e. UNet, cell segmentation was further developed. The network is of the encoder-decoder structure, its input is first be downsampled and then upsampled. In the upsampling process, low-layer features corresponding to the downsampling layer are connected to the corresponding upsampling layer. The network gets high-level semantics without losing much low-level information and works well

with only small amount of training data. Zhou et al. [12] redesigned the skip-connection of UNet and improved segmentation performance by reducing semantic differences of feature maps in encoder and decoder subnetworks.

Payer et al. [5] integrated ConvGRU into a stacked hourglass network for instance segmentation and tracking of cells. ConvGRU not only extracts local features, but also memorizes inter-frame information [13]. The stacked hourglass network is similar to UNet, its input is first be downsampled and then upsampled [14]. This integrated structure can perform cell segmentation well even in the case of a very close membrane. However, when cells are small, it does not perform well. Arbelle et al. [15] proposed a network structure for inter-frame segmentation combining ConvLSTM and UNet. Like ConvGRU, ConvLSTM has spatio-temporal characteristics [16]. The integrated structure can perform excellent segmentation even in the case of partial disappearance of cells, but does not perform well in the case of less training data. The work of Payer et al. [5] and Arbelle et al. [15] are similar.

As long as cells are accurately detected or segmented, tracking will be largely simplified. In Ref. [5], although only the intersection over union is used for inter-frame cell associations, it achieves desired tracking performance.
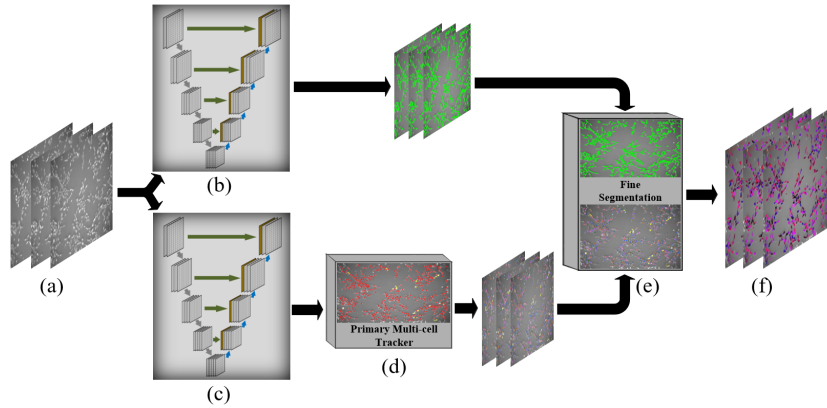

Fig. 1. Overview of our proposed tracking framework. (a) Input. (b) UNet for primary cell segmentation. (c) UNet for cell centroid detection with multi-frame images. (d) Primary multi-cell tracker. (e) Fine segmentation. (f) Final tracking results.

3 Method

In this section our proposed method is detailed. As shown in Fig. 1, it has four portions: cell centroid detection with multi-frame images, primary multi-cell tracker, primary cell segmentation, fine segmentation.

3.1   Cell Centroid Detection with Multi-frame Images

To identify cells in highly dense population, it is a useful technique to identify cell centroid first. Here UNet [6] (UNet-DET) is used to locate cell centroid.

Mitotic cells are defined as those cells before, during and after mitosis. During mitosis, obvious morphological changes usually occur, which make them look different from normal cells, i.e., cells in non-mitosis status. In Fig. 2, from left to right, morphological changes of a cell before, during and after mitosis.
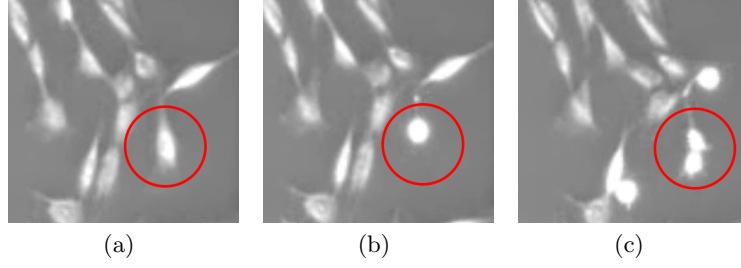

(a)   (b)   (c)

Fig. 2. Morphological changes in mitosis.

Pixels are categorized into three categories: mitotic cells, normal cells and backgrounds. If information in previous nearby frames is included, network can more accurately learn to identify mitotic cells [17].

Different from usual single-frame input method, we feed incorporative consecutive pre-N_{input} frames into the network. This approach does improve cell centroid detection a lot. Combined with past image information, the network can extract living cell behavioral features, and then screen out impurities that do not change shape.

Compared with Ref. [5, 15], the overall network complexity based on UNet-DET does not increase much. Only the number of weight parameters in the first layer of the network increases.

A cross-entropy loss function is used to train the network. Mitotic cells are few, and more attention should be paid to them by setting a weight map. The network loss function is defined as in Ref. [6]:

$$L = -\frac{1}{T}\sum_{i=1}^{T} w(i)\log \frac{\exp(h(i,g(i)))}{\sum\limits_{j=1}^{C}\exp(h(i,j))} \tag{1}$$

where T denotes the total number of pixels. $w(\cdot)$ is a weight map. $g(\cdot)$ is the real category corresponding to the input pixel. $h(i,j)$ denotes the final output of category j when input pixel i.

Thresholding inference results, using the flood-fill algorithm [18] to fill internal holes of connected area. Extracting contour of connected area, computing its

centroid as cell centroid. High-precision cell centroid detection and classification are acquired without increasing much network overhead.

3.2 Primary Multi-cell Tracker

Different from common object tracking, cell lineage needs to be built up during tracking. If a cell is categorized as a mitotic cell in multiple consecutive frames, it is highly likely to undergo mitosis. A mitosis detection algorithm, using a local cell status matrix, is proposed. The matrix is created in the beginning of new trajectory. Elements of the matrix record the status of a cell: the centroid position (X, Y), its sequence number (Z), whether it is a normal or mitotic cell. When the number of mitotic cells is larger than a given threshold, it can be concluded that mitosis occurs.

As long as images are captured in high frame rate and cells are precisely detected, it is possible to use overlap intersection-over-union (IOU) to build inter-frame object associations [8, 19], and then ideal tracking is performed.

When the cell centroid is located, a bounding box with a size of $N_{size} \times N_{size}$ is created around the centroid, where N_{size} is the average length of cells scaled in number of pixels. With the assumption that sequences have the high enough frame rate, inter-frame cells can be associated by only using the IOU of bounding boxes. Each newly detected cell is needed to be associated with an existing trajectory. The association strategy is computed as:

$$\Theta(t, D) = \mathop{arg\,ma}\limits_{d \in D} x(\Lambda(d, t)) \tag{2}$$

$$d(t, D) = \begin{cases} \Theta(t, D), & \Lambda(\Theta(t, D), t) \geq \alpha \\ N, & \Lambda(\Theta(t, D), t) < \alpha \end{cases} \tag{3}$$

where t denotes the cell at the end of the trajectory to be associated with, D the set of candidate detected cells in the current frame, $\Lambda(\cdot, \cdot)$ the IOU of both. α is the minimum overlapping intersection that allows to associate. N denotes that there are no associating candidate cells.

If the largest IOU that is less than α, the candidate cell will be discarded.

The trajectory is terminated when it has no associated cells. For a candidate cell that has no existing trajectory to associate, a temporary new trajectory is created. Short trajectories are most likely pseudo trajectories caused by impurity interference, and therefore are discarded.

If a mitosis event is detected, the original trajectory is terminated. At the same time, cell lineage is established between the mitotic mother cell and its two newborn daughter cells. This distinguishes normal cells which newly enter the field of view from newborn daughter cells.

Primary tracking results of cell centroid detection are acquired, cell IDs of the same trajectory are same.

3.3   Primary Cell Segmentation

UNet [6] (UNet-SEG) is used for primary segmentation of cells. During this stage, image pixels are categorized into cell boundaries, cell interiors, and backgrounds. The cross-entropy loss function is used to train the network. More attention is paid on cell boundaries by setting a weight map.

Threshold is performed to acquire cell segmentation by using inference results of cell boundaries and cell interiors. There may be holes inside segmented cells and the flood-fill algorithm [18] is used to fill these internal holes.

When many cells are close each other, each cell is not separated at this stage. When primary segmentation is finished, as shown in Fig. 3 (b), cells close each other may be segmented as a blob, a piece of connected area.
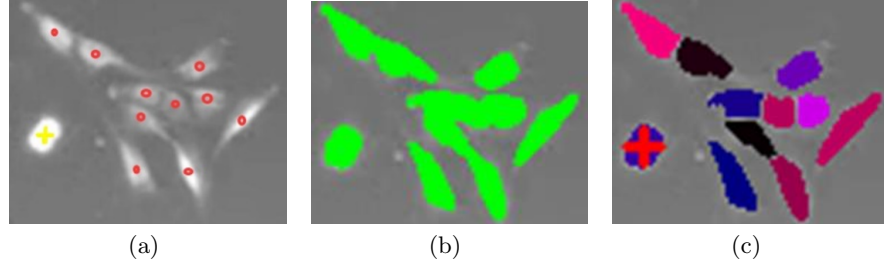

| (a) | (b) | (c) |

Fig. 3. Dense cell segmentation results. Cross: mitotic cells. Dot: normal cells. (a) Original image and cell centroid detection results. (b) Primary cell segmentation results. (c) Fine segmentation results.

3.4   Fine Segmentation

Results from primary segmentation may contain many connected area as shown in Fig. 3 (b). When cells are dense, multiple cells are segmented together. In this section, fine segmentation is conducted to separate each cell individually, which jointly use primary tracking results of cell centroid detection from Sect. 3.2 and primary cell segmentation results from Sect. 3.3.

Assuming cell boundary is closest to its centroid for non-overlapping small cells of similar size, each pixel in a connected area is assigned to a cell centroid contained in this connected area according to the following formulation:

$$P(p_{pixel}) = \arg\min_{p \in P_{\mathrm{det}}}(d(p_{pixel}, p)) \tag{4}$$

where P_{det} is a set of cell centroid contained in a connected area, and p_{pixel} denotes pixels in the connected area, $d(\cdot, \cdot)$ denotes the Euclidean distance.

However, it would be a very time-consuming task to calculate Euclidean distance from each pixel to each cell centroid. Therefore, Voronoi [20] is used to

accelerate pixel assignment. Pixel assignment is shown as Algorithm 1. Fig. 3 (c) shows fine segmentation results.

Algorithm 1 Fine Segmentation

Input:
  Primary cell segmentation $SEG = SEG_1, SEG_2, \ldots, SEG_M$;
  Primary tracking results of cell centroid detection $P_{det} = P_1, P_2, \ldots, P_N$;
Output:
  Fine segmentation result SEG_RES;
1: Initialize a zero matrix SEG_RES;
2: Calculate the Voronoi diagram according to P_{det};
3: Label the connected domain of Voronoi diagram, acquire LAB_VOR;
4: **for** seg in SEG **do**
5:     P_{inside} = cell centroid contained by seg;
6:     **if** len(P_{inside})==1 **then**
7:         SEG_RES += seg;
8:     **else**
9:         **if** len(P_{inside})>1 **then**
10:             **for** p in P_{inside} **do**
11:                 SEG_RES += $seg \bigcap LAB\_VOR(p)$;
12:             **end for**
13:         **end if**
14:     **end if**
15: **end for**

---

4   Experiment

The effectiveness of our method is evaluated with datasets of Cell Tracking Challenge [9]. The dataset for each kind of cells provides two training sequences with GT (ground truth) and two test sequences without GT. GT includes TRA (tracking) GT and SEG (segmentation) GT. TRA GT essentially contains cell centroid of all sequences. SEG GT is few, which makes cell segmentation more difficult.

4.1   Training

UNet-DET and UNet-SEG both are composed of 5 downsampling layers and 5 upsampling layers. Adam optimizer [21] is used and the learning rate is set as *0.001*. The exponential decay rate of learning rate is set as *0.95*, the global step size of the decay is set as *4* times of the number of sample set. A weighted cross-entropy loss function is used for training to pay more attention on mitotic cells or cell boundaries. To augment samples, we apply horizontal and vertical flips, and randomly add Gaussian noise or salt and pepper noise on sample set.

For UNet-DET, incorporative consecutive pre-N_{input} frames are fed into the network. The dimension of input is (H, W, N_{input}), here N_{input} = 3. The label is TRA GT of last frame. We have modified TRA GT. Mitotic cells are defined as cells of $N_{mitisis}$ frames before and after mitosis, here $N_{mitosis}$ = 2. Weight of each category is set as: 0.5-mitotic cells, 0.3-normal cells, 0.2-backgrounds.

For UNet-SEG, SEG GT are categorized into cell boundaries and cell interiors. When the amount of SEG GT is scarce, original images and SEG GT are cropped centered on each cell centroid. N_{cell} training samples with a size of $S_{crop} \times S_{ceop}$ are cropped from each original image. Here S_{crop} is set as 5 times of the mean size of cells, N_{cell} denotes the number of cells in the image. Weight of each category is set as: 0.5-cell boundaries, 0.3-cell interiors, 0.2-backgrounds.

4.2   Comparison of Multi-frame Input and Single-frame Input

The advantage of multi-frame input over single-frame input is evaluated with Cell Tracking Challenge datasets PhC-PSC and Fluo-HeLa [7], which have more mitosis. The first half of the two training sequences are used as training samples and the other half are used as test samples. Performance of cell centroid detection is scaled with three metrics, i.e., Precision, Recall, F1-score. These metrics on normal cells and mitotic cells are shown in Table 1 respectively.

From Table 1, three metrics are improved 15% with mitotic cells, while slightly improved with normal cells. Inter-frame morphological changes of mitotic cells are obvious, retaining the historical information can improve detection performance.

In our multi-cell tracking algorithm, the improved mitotic cells detection performance will improve the mitosis event detection performance. Table 2 shows the mitosis event detection performance compared with different input modes. Due to our mitosis detection algorithm is strict to determine whether mitosis has occurred, Precision is high of both. The other two metrics of multi-frame input are improved about 20%. Therefore, the reconstruction of the cell lineage will be improved.

Table 1. Detection performance of normal cells and mitotic cells

Status	Input Frame	Precision (%)	Recall (%)	F1 score (%)
Mitotic cells	1	44.3	49.7	46.8
	3	**60.1**	**65.3**	**62.6**
Normal cells	1	93.2	91.6	92.4
	3	**94.3**	**93.7**	**94.0**

4.3   Evaluations on Cell Tracking Benchmark

Cell tracking performance is evaluated in Cell Tracking Benchmark with 2D datasets of Cell Tracking Challenge [7]. Performance of tracking is scaled in

Table 2. Detection performance of mitosis event

Dataset	Input Frame	Precision (%)	Recall (%)	F1 score (%)
Phc-PSC	1	**87.6**	26.8	41.1
	3	86.0	**47.8**	**61.5**
Fluo-Hela	1	77.3	37.2	50.2
	3	**87.7**	**63.6**	**73.7**

TRA (tracking accuracy), SEG (segmentation accuracy), and OP_{CTB} (the mean of both) [7].

Based on the ranking announced on the day April 30th, 2019, which can be seen at the web of Cell Tracking Challenge [9], performance metrics of our proposed method are listed in Table 3. For each dataset, generally performance metrics of more than 20 methods are ranked [9], including the original UNet (FR-Ro-GE) [6], the globally trained UNet (FR-Fa-GE) [22], the method of ConvLSTM integrated into UNet (BGU-IL) [15], the method of ConvGRU integrated into the stacked hourglass network (TUG-AT) [5], and the method of global threshold and global associations with spatio-temporal (KTH-SE) [9].

Table 3. Quantitative Comparison of Cell Tracking Benchmark (%)

		Phc-PSC	Fluo-Hela	Fluo-SIM+	Fluo-GOWT1	
OP_{CTB}	1st	80.4	95.3	88.2	95.1	
	2nd	80.4	94.4	88.1	93.4	Ours
	3rd	80.1	94.2	87.8	92.3	KTH-SE(1-4)
				7th 86.0	12th 87.0	HD-Hau-GE
SEG	1st	68.2	91.9	80.7	92.7	CVUT-CZ
	2nd	66.5	90.3	80.2	92.1	RWTH-GE
	3rd	65.3	90.2	79.2	89.4	FR-Ro-GE
				4th 78.4	16th 79.0	FR-Fa-GE
TRA	1st	95.9	99.1	97.5	97.9	BGU-IL(1-4)
	2nd	95.0	99.1	97.3	97.6	TUG-AT
	3rd	94.3	98.8	96.6	96.7	
			4th 98.7	9th 93.6	4th 94.9	

Experiments show excellence of our method on datasets Phc-PSC and Fluo-Hela, highly dense cell datasets. Our method achieves new state-of-the-art performance on SEG and OP_{CTB} of dataset Fluo-Hela. 2nd on TRA of dataset Phc-PSC is achieved. Though these two datasets have very few SEG GT, our method performs excellent.

Our method does not perform very well on datasets Fluo-SIM+ and Fluo-GOWT1. A possible solution to this problem is fully taking advantage of image information when making primary cell segmentation split.

Fig. 4 shows multi-cell tracking performance of our method on multiple datasets. For the consideration of clarity, only a portion of field of view is selected and enlarged. Different kind of cells have different morphology. We track trajectories of cells and get each cell segmentation. Fine segmentation results on highly dense cell population is shown as in Fig. 4 (a). As shown in Fig. 4 (c), cells can be segmented accurately even when partly disappeared. Cells can be segmented accurately when their gray level is similar to that of background.

Fig. 5 shows cell spatio-temporal trajectories of our method on dataset Phc-PSC. It shows trajectories of all cells, as well as evolution of cell lineage. Cells will undergo mitosis over time or leave the field of view. As the frame number increases, cell trajectories become dense.
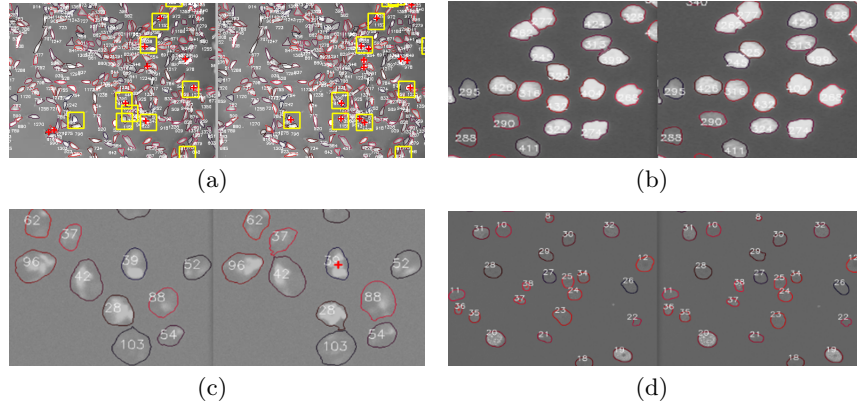

(a)	(b)
(c)	(d)

Fig. 4. Cell tracking results. For each pair of images, the left one is the previous frame, and the right one is the current frame. White numbers: trajectory IDs. Yellow boxes: detected mitosis. Red crosses: mitotic cells. Datasets: (a) Phc-PSC. (b) Fluo-Hela. (c) Fluo-SIM+. (d) Fluo-GOWT1.

5   Conclusion

We propose a multi-cell tracking framework, which jointly use detection and segmentation. Cell centroid detection is conducted using a UNet with multi-frame input images. Detection of mitotic cells is improved without increasing much network overhead, and therefore improve the detection performance of mitosis event by our mitosis detection algorithm. With our method, normal cells newly entering the field of view can be distinguished from newborn daughter cells. Another UNet is utilized to acquire primary cell segmentation. Fine segmentation is conducted to separate each cell individually, which jointly use primary tracking results of cell centroid detection and primary cell segmentation results.
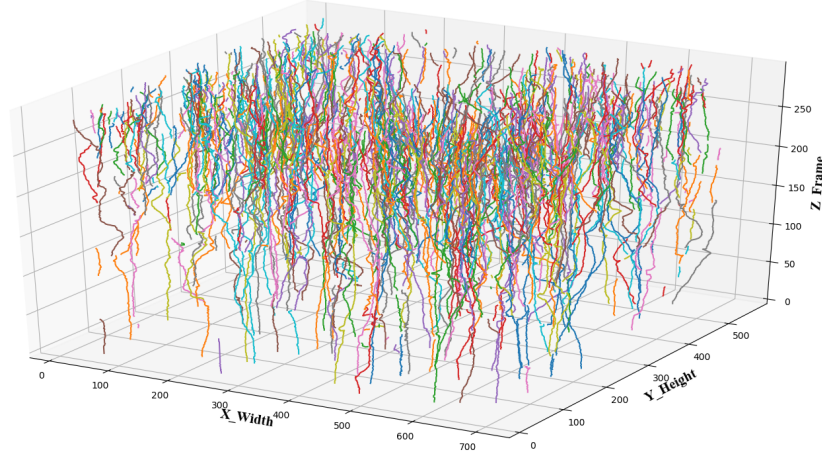
Fig. 5. Cell spatio-temporal trajectories of Phc-PSC.

Evaluations are conducted to compare our method with other methods with datasets in Cell Tracking Challenge. Due to jointly use detection and segmentation, our method performs excellent and achieves a new state-of-the-art performance on dataset Fluo-Hela.

Performance on some datasets is still not very ideal. In future works, fine segmentation will be further optimized, and more image information will be used for more accurately segmentation and tracking.

References

1. Ren Y, Xu B, Zhang J, et al. A generalized data association approach for cell tracking in high-density population[C]//2015 International Conference on Control, Automation and Information Sciences (ICCAIS). IEEE, 2015: 502-507.
2. He T, Mao H, Guo J, et al. Cell tracking using deep neural networks with multi-task learning[J]. Image and Vision Computing, 2017, 60: 142-153.
3. Yang F, Mackey M A, Ianzini F, et al. Cell Segmentation, Tracking, and Mitosis Detection Using Temporal Context[J]. Lecture Notes in Computer Science (LNCS), 2005, 8(Pt 1):302-309.
4. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2016: 770-778.
5. Payer C, Štern D, Neff T, et al. Instance segmentation and tracking with cosine embeddings and recurrent hourglass networks[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, Cham, 2018: 3-11.
6. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]// International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, Cham, 2015: 234-241.

7. Maška M, Ulman V, Svoboda D, et al. A benchmark for comparison of cell tracking algorithms[J]. Bioinformatics, 2014, 30(11): 1609-1617.
8. Bochinski E, Eiselein V, Sikora T. High-speed tracking-by-detection without using image information[C]//2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2017: 1-6.
9. Ulman V, Maška M, Magnusson K E G, et al. An objective comparison of cell-tracking algorithms[J]. Nature methods, 2017, 14(12): 1141.
10. Luo W, Xing J, Milan A, et al. Multiple object tracking: A literature review[J]. arXiv preprint arXiv:1409.7618v4, 2017.
11. Ciresan D, Giusti A, Gambardella L M, et al. Deep neural networks segment neuronal membranes in electron microscopy images[C]//Advances in neural information processing systems. 2012: 2843-2851.
12. Zhou Z, Siddiquee M M R, Tajbakhsh N, et al. Unet++: A nested u-net architecture for medical image segmentation[M]//Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA). Springer, Cham, 2018: 3-11.
13. Ballas N, Yao L, Pal C, et al. Delving deeper into convolutional networks for learning video representations[J]. arXiv preprint arXiv:1511.06432, 2015.
14. Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]//European Conference on Computer Vision (ECCV). Springer, Cham, 2016: 483-499.
15. Arbelle A, Raviv T R. Microscopy Cell Segmentation via Convolutional LSTM Networks[J]. arXiv preprint arXiv:1805.11247, 2018.
16. Xingjian S H I, Chen Z, Wang H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[C]//Advances in neural information processing systems. 2015: 802-810.
17. Sadeghian A, Alahi A, Savarese S. Tracking the untrackable: Learning to track multiple cues with long-term dependencies[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017: 300-311.
18. Khudeev R. A new flood-fill algorithm for closed contour[C]//2005 Siberian Conference on Control and Communications. IEEE, 2005: 172-176.
19. Bochinski E, Senst T, Sikora T. Extending IOU based multi-object tracking by visual information[C]//2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2018: 1-6.
20. Okabe A, Boots B, Sugihara K, et al. Spatial tessellations: concepts and applications of Voronoi diagrams[M]. John Wiley & Sons, 2009.
21. Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
22. Falk T, Mai D, Bensch R, et al. U-Net: deep learning for cell counting, detection, and morphometry[J]. Nature methods, 2019, 16(1): 67.