



Cell tracking with multifeature fusion

Fei Wang¹ · Huawei Li¹ · Wanli Yang¹ · Shidong Jin¹ · Peng Gao²

Accepted: 5 May 2023 / Published online: 6 June 2023
© The Author(s) 2023

Abstract

Cell tracking is currently a powerful tool in a variety of biomedical research topics. Most cell tracking algorithms follow the tracking by detection paradigm. Detection is critical for subsequent tracking. Unfortunately, very accurate detection is not easy due to many factors like densely populated, low contrast, and possible impurities included. Keeping tracking multiple cells across frames suffers many difficulties, as cells may have similar appearance, they may change their shapes, and nearby cells may interact each other. In this paper, we propose a unified tracking-by-detection framework, where a powerful detector AttentionUnet++, a multimodal extension of the Efficient Convolution Operators algorithm, and an effective data association algorithm are included. Experiments show that the proposed algorithm can outperform many existing cell tracking algorithms.

Keywords Cell tracking · Data association · Detection

1 Introduction

Benefiting from microscopy imaging technologies and computer algorithms, scientists can follow moving cell behavior and construct cellular lineages by processing digital images. This knowledge is critical in biological research, drug development, medicinal or preventive therapies[3]. For example, cell tracking can help developmental scientists to understand the developmental history of cell functions [1] identify the role of differential gene expression in directing cell fate, analyze mutant defects, and verify development models. In assistive reproduction

✉ Peng Gao
pgao@qfnu.edu.cn

Fei Wang
wangfeiz@hit.edu.cn

¹ School of Electronic and Information Engineering, Harbin Institute of Technology, Shenzhen, Pingshan No.1 Road, Shenzhen 518000, Guangdong, China

² School of Cyber Science and Engineering, Qufu Normal University, Jingxuanxi Road, Qufu 273165, Shandong, China

labs, cell tracking can help to evaluate spermatozoid motility before artificial insemination [25]. Cell tracking is broadly used to evaluate spermatozoid motility in assistive reproduction technology [7]. By tracking single cell movement, Winter et al. [26], evaluate the role of deficiencies in axonal organelle transport in the pathogenesis of neurodegenerative diseases. Kwak YH et al. [15], studied cell chemotaxis by tracking a single cell's migration route on a microfluid chip under culture conditions. Yue et al. [27], recorded long-term live images of developing mouse hearts, by tracking cells and reconstructing cell lineages, revealing the ventricle chamber formation.

Most discriminative algorithms follow the tracking-by-detection paradigm [7, 15, 26, 27], which treats the tracking task as a detection problem. They employ a classifier or a regressor to process both target and background representations. The detector is critical to the tracking performance. Unfortunately, because of possible impurities may coexist with cells, or the contrast between the target and the background is not high, the detector may not success all the time.

On the other hand, single object tracker can partially address this problem, as it does not rely on detection very much. Once the target's initial position is assigned, SOT (Single Object Tracking) can capture the target in the following frames, and even the target disappears for a short time, without needing to detect the target in each frame. However, after certain frames, the SOT may drift.

However, these SOT algorithms usually suffer from the data imbalance issue between positive and negative samples for online model updating. In the search window of a tracker, positive samples are usually around the target center, while negative samples occupy the other area. This imbalance may cause inefficient training, and cause the model apt to drift. This problem is exacerbated in the case of multicell tracking tasks. As MOT (Multiple Object Tracking) suffers frequent interactions between cells with similar appearances, the tracker is more prone to drift. Thus, it is imperative to focus on a small number of hard examples during online updating to alleviate the drifting problems.

Recently, Discriminative Correlation Filter (DCF)-based SOT algorithms, esp. those combined with deep CNN (convolutional neural networks), have achieved surprising performance. However, they suffer problems of heavy computation overhead and the risk of severe overfitting. ECO incorporates a factorized convolution operator and a generative model of the training sample distribution to reduce computation complexity. Unlike many other DCF trackers, which update their models in each frame, ECO utilizes a sparser updating scheme to combat model drifting. Therefore we use ECO (Efficient Convolution Operators) as the base SOT for our multiple cell tracker.

To address the above problems, we propose a framework named as ECOMA (Efficient Convolution Operators with Multi-feature and Association). Our contributions can be summarized as follows:

- (1) An accurate detector named as AttentionUnet++ is proposed, where spatial-temporal attention mechanism is introduced. It can adapt to more broad types of cells under varying conditions. Compared with Unet++, it is more robust.

- (2) An ECO-based SOT is extended into a MOT tracker, where multimodal feature fusion and data association are included. In the data association module, a network structure is proposed to evaluate cell similarity across frames.

This paper is organized as follows. Section 2 introduces the AttentionUNet++ detector. Section 3 introduces the ECOMA tracking algorithm and the SiameseNet-like data association algorithm. Section 4 presents some experimental results to verify our algorithms.

2 Related works

Tracking-by-detection, is currently the dominant paradigm of multitarget tracking, where detection is critical for successful tracking. Unfortunately, very accurate detection is not easy due to many factors like densely populated, low contrast, and possible impurities included, false detection may occur, which may lead to poor performance in the following tracking. Another challenging in multicell tracking comes from the interaction between nearby cells. As cells may have similar appearance and may change their shapes, it is not easy to associate correctly across different frames. Therefore, data association is very critical. In this section, we give a brief overview on related works from two perspectives, i.e., detection and tracking.

Unet is widely used in medical image processing [21]. Its encoding-decoding structure endows it with distinguish segmentation performance even when there is not much training data. Unfortunately, for different data set, it has different optimal depth. In order to obtain the best segmentation effect, it may be necessary to try multiple Unets with different depths. Unet++ breaks the constraints of Unet [28].

DESU-US takes advantage a cell detection algorithm based the combination of motion diffusion-based partial differential equation (PDE) and active contours [5]. It also produces a tracking algorithm adopting a variational joint local-global optical flow technique to determine the motion vector field. Cell motion was predicted using a maximum likelihood criterion, which considers motion vectors and spatial cell features jointly.

DREX-US is a graph-based method for multicell target tracking [26]. It reduces the error rate and implementation complexity compared to approaches based on bipartite matching. Data association was solved using a graph-based cost approximating a posteriori probability across a window of future detection data.

As cell detection and data association are most frequently two separate stages, there is no guarantee of preserving coherence between these two stages. This separation may adversely affect tracking performance. To address this problem, Hayashida proposes an MPM-Net, which jointly represents both detection and association [15], where motion and position map (MPM) represents both the motion vector and the position likelihood map.

Li et al. [17], propose a cell tracking method which jointly uses a dynamic memory network and template matching. Cell detection is implemented by a fully convolutional neural network, and cell tracking is implemented with multiple dynamic

memory units. The template is dynamically updated using an attention LSTM to cope with changing cell appearance. Cell motion statistics based on motion constraint are conducted to improve the robustness. Experimental results demonstrate its good performance.

He et al. [13], proposed a novel cell tracking method is proposed by using CNNs as well as multitask learning (MTL) techniques. After initializing the cell positions in the first frame, the particle filter model is applied to produce a set of possible bounding boxes in the subsequent frames. The multitask learning observation CNN extracts robust cell features, and evaluates those candidate bounding boxes, with the one with the highest confidence probabilities selected as the final predicted position.

Zhou et al. [29], jointly used two UNets for cell detection and segmentation respectively. The one for cell detection extracts both interframe and intraframe spatiotemporal information. The other one can perform cell segmentation as well as mitosis at the same time. The algorithm works well in high densely populated cell images.

3 The proposed ECOMA tracker

As mentioned before, successful tracking of multiple cells depends on both powerful detection and tracking techniques. In this paper, we propose a unified frame work from the two perspectives for multicell tracking, as shown in Fig. 1.

At each frame, target detection (DET) is conducted with a robust detector AttentionUNet++, which incorporates both spatial and temporal attention mechanisms into Unet++. Then ECOMA is activated to track each newly detected target, which

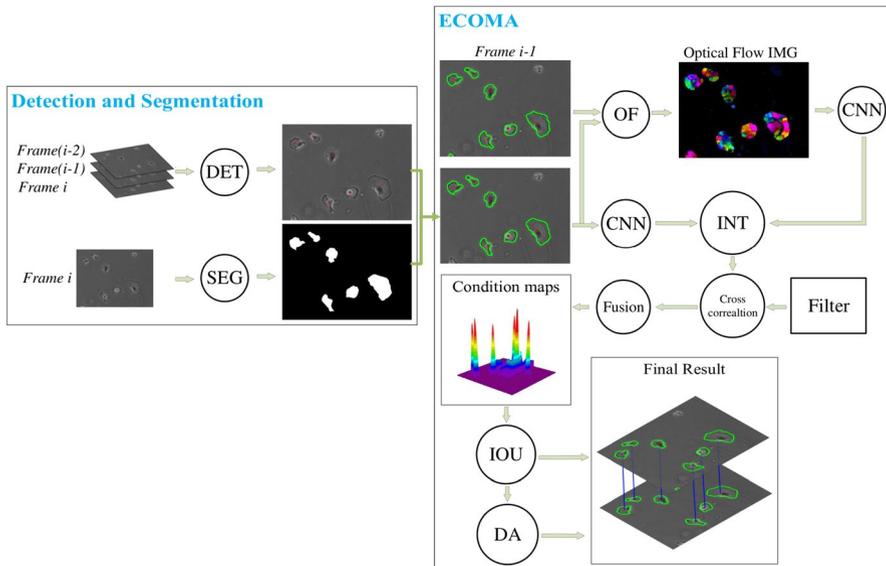


Fig. 1 Framework of the proposed cell tracking algorithm

takes both advantage of both deep appearance features and deep motion features. Deep appearance features are extracted with a deep CNN. Optical flow maps [4] are constructed from two nearby frames, and then deep motion features are extracted from optical flow maps using another deep CNN. In order to balance computation overhead and feature extraction performance, we use the output features of bottleneck block 12 and 17 of MobileNet [23] as features at different depths.

For those with a low tracking score, marked as “lost” target, the ECOMA tracker is suspended and data association is activated to evaluate the similarity between the lost target and those “tracked” targets in the previous frame. Once the lost target is linked to a tracklet through data association, it is updated as “tracked”, and the tracking process is restored. For any new detection, which cannot be linked to a tracklet, a new trajectory is initiated. This data association algorithm does not have to be conducted for each cell in each frame. It is activated only when tracking is unreliable. Experiments show that the proposed algorithm can outperform many existing cell tracking algorithms.

3.1 Cell segmentation and detection

In Fig. 2, if the attention module is bypassed and removed, then it becomes a Unet++. The overall architecture of Unet++ is kind of similar to that of Unet, both have downsampling networks (encoders), upsampling networks (decoders) and intermediate hopping connections. The down arrow in the figure denotes downsampling, the up arrow upsampling, and the dotted arrow hop connection. $X_{i,j}$ in the circle represents convolution layer, and L represents the loss function. Unet++ combines Unet networks with different depth. These UNet networks share the same encoding network, and their decoders are intertwined.

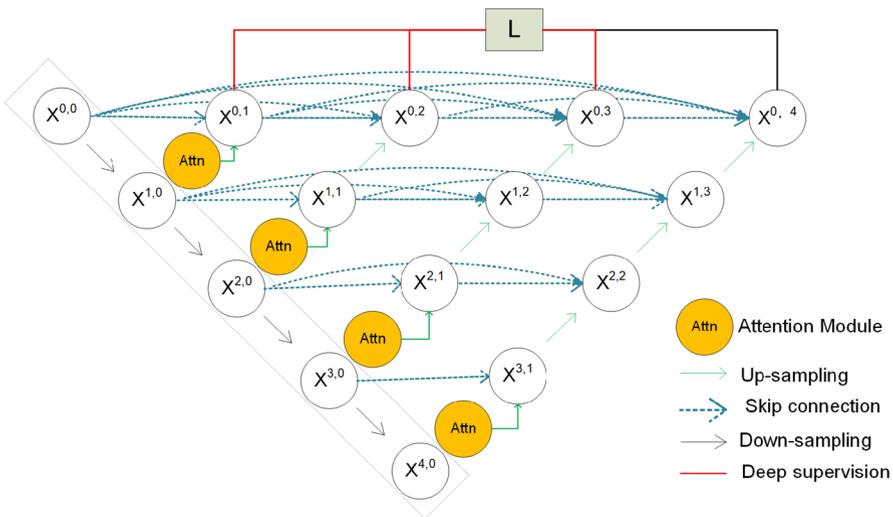


Fig. 2 Network structure of attentionUnet++

deep supervision, this design can speedup reasoning with high accuracy. For each node $X_{i,j}$, i is the index along the sampling layer, and j is the index along the hopping connection. Let $x_{i,j}$ represent the output feature map of node $X_{i,j}$, which can be calculated as:

$$x^{i,j} = \begin{cases} H(x^{(i-1,j)}) & \text{if } j = 0 \\ H([\![x^{i,k}]_{k=0}^{j-1}, u(x^{i+1,j-1})\!\]) & \text{if } j > 0 \end{cases} \quad (1)$$

where $H(\cdot)$ represents the activation function, and $u(\cdot)$ represents the upsampling and $[\cdot]$ represents the connection. The node with $j = 0$ only receives the input from the upper layer of the encoder; the node with $j = 1$ receives two inputs from the adjacent encoder network layer. The node with $j > 1$ receives $j + 1$ inputs, of which j inputs are from the outputs of the first j nodes on the same hopping connection, and the last input is from the node just below it.

In the backbone, the convolution block composed of multiple convolution layers, can accumulate all the previous feature maps and deliver them to the last node. Hopping connections combine the shallow feature map from the encoder network with the deep feature map from the decoder, which can effectively restore the fine-grained features of the object, and help the network recover the lost image information and improve the performance.

3.1.1 AttentionUNet++

It is expected that the network can extract spatiotemporal information. Multiple frames are connected along the channel dimension, for example, the original image frame dimension is hWC . When combining the images of the past n frames, the data sent to the network have a dimension of $hW(C(1+n))$. This method only allows the network to receive multi-frame images by changing the input data dimension, which is used to improve the detection performance without adding the parameters of the input layer. Attention mechanism can emphasize important features and suppress non-important features. Since the input tensor to our UNet++ has a dimension of $hW(C(1+n))$, the different channels of the input tensor are the image information of different frames, and convolution operations extract features by mixing the cross channel and spatial information.

Inspired by [11], the channel attention map and spatial attention map are applied to feature maps in turn. Each branch can learn important information along the channel and spatial dimension, respectively. Therefore, the network will be enhanced to focus on targets. The network structure of the temporal attention and spatial attention mechanisms are shown in Figs. 3 and 4.

The structure of the temporal attention module is shown in Fig. 3 or calculated using Eq (2), and features across channels are utilized to generate an attention map. Since the pooling operations and shared MLP (multilayer perception) are applied on each channel of the input feature maps across $n + 1$ nearby frames, it makes sense to

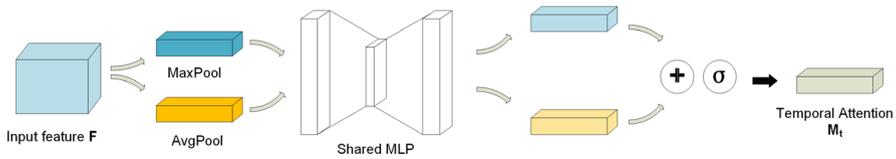


Fig. 3 Temporal attention

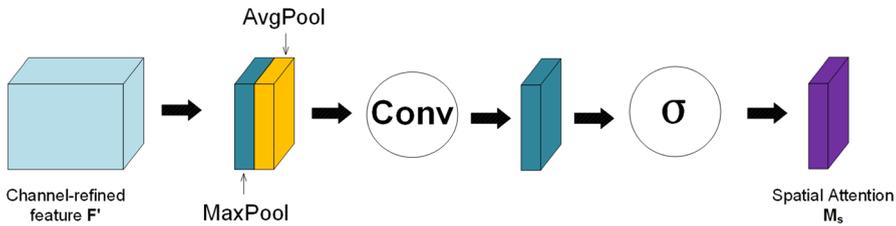


Fig. 4 Spatial attention

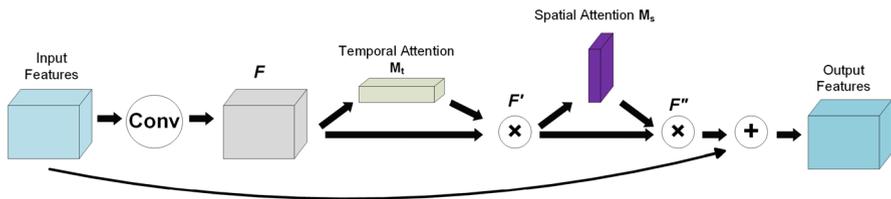


Fig. 5 Spatial-temporal attention

focus on “which frames.” In order to effectively calculate channel attention maps, we compress the spatial dimension of the input feature maps.

$$M_t(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \tag{2}$$

$$M_s(F) = \sigma(f^{7 \times 7}[\text{AvgPoll}(F); \text{MaxPool}(F)]) \tag{3}$$

Different from channel attention, spatial attention focuses on “which location,” and works as a complement to channel attention. As calculated in Eq. (3), average pooling and maximum pooling operations are very efficient, and they are applied along the channel dimension. The two corresponding feature maps $F_{\text{avg}}^S \in R^{H \times W \times 1}$ and $F_{\text{max}}^S \in R^{H \times W \times 1}$ are concatenated, and then convoluted with a kernel of 7×7 , and finally normalized with a sigmoid function, and corresponding structure is shown in Fig. 4.

A spatial-temporal attention module is constructed as shown in Fig. 5. Attention (weight) maps from Figs. 3 and 4 are applied onto feature maps by multiplications. When multiple such modules are integrated into Unet++, therefore

AttentionUNet++ is built up as shown in Fig. 2, where L represents the loss function. Using these embedded attention modules, let the network emphasize the channel in temporal and spatial features. Attention Unet++ takes multiple neighboring frames as inputs, rather than single frame. Multiframe image information is spliced on the channel, therefore detection can be performed with higher accuracy.

3.2 ECO tracker

Here ECO is briefly summarized on which our multicell tracker is built [9]. Suppose totally there are D channels of appearance or motion feature maps, $\phi^1, \phi^2, \dots, \phi^D$, and each ϕ^d has a size of N^d . As they have different size, before they are combined, they are interpolated into continuous domain:

$$\Phi\{\phi^d\}(t) = \sum_{n=0}^{N^d-1} \phi^d[n]c_d\left(t - \frac{T}{N^d}n\right) \tag{4}$$

where $\Phi\{\phi^d\}(t)$ is $\phi^d(n)$'s counterpart in continuous domain. c_d is a periodic function as expressed in the following equation:

$$c_d(t) = \sum_{-\infty}^{+\infty} b_d\left(\frac{N^d}{T}\left(t - \frac{nT}{2N^d}\right)\right) \tag{5}$$

where b_d is an interpolation kernel, e.g., a cubic spline. $b_d(t)$ for channel d is scaled to the sampling frequency T/N^d

Conducting convolution on $\Phi(\phi^d)$ with a filter f^d in continuous domain creates a confidence map of channel d . Summarizing all those confidence maps on all channels creates the combined confidence map:

$$S_f(x) = \sum_d^D f^d * \Phi(\phi^d) \tag{6}$$

The target can be localized by searching the peak value across the combined confidence map [10]. The tag y_j is also defined on continuous domain, which denotes the expected feature map when f^d is applied on ϕ^d . All these filters can be learned by minimizing the following loss function:

$$E(f) = \arg \min_f \sum_{j=1}^m \alpha_j \left\| S_f\{x_j\} - y_j \right\|^2 + \sum_d^D \left\| w f^d \right\|^2 \tag{7}$$

where m denotes the total number of samples, and α_j controls the weight for each sample, and the second item denotes the regularizing penalty.

Applying Fourier Transform to (5)

$$\widehat{S}_f\{x\}[k] = \sum_d^D \widehat{f}^d[k] \Phi^d[k] \widehat{c}_d[k] \tag{8}$$

where hat functions corresponds to the Fourier transform of corresponding items in (6). It is very time consuming to minimize the function. According to Parseval’s formula, Eq. (6) is equivalent to resolving the following formulation in Fourier domain [8]:

$$E(f) = \arg \min_f \sum_{j=1}^m \alpha_j \left\| \sum_d^D \widehat{f}^d \Phi_j^d \widehat{c}_d - \widehat{y}_j \right\|^2 + \sum_d^D \left\| \widehat{w} * \widehat{f}^d \right\|^2 \tag{9}$$

To minimizing Eq. (9) is equivalent to resolving the following equation:

$$(A^H \Gamma A + W^H W) \widehat{\mathbf{f}} = A^H \Gamma \widehat{\mathbf{y}} \tag{10}$$

where $\widehat{\mathbf{f}}^d$ and $\widehat{\mathbf{y}}$ are vectorizations of the Fourier coefficients of f^d and y_j . H denotes the conjugate transpose. A denotes a matrix with diagonal blocks, and each block contains elements of the form $\Phi_j^d \widehat{c}_d[k]$. Γ is a diagonal matrix of weights α_j , and W corresponds is a convolution matrix with the kernel $\widehat{w}[k]$. To localize the target, the peak value on the inverse Fourier Transform of the response map is searched.

Unfortunately, D is often a very large number. To speedup calculation, ECO algorithm compresses the D filters into smaller $C(C < D)$ filters. To avoid the number of samples being intractable, ECO takes advantage of Gaussian Mixture Model to build a generative distribution model of samples.

3.3 Data association

Using the ECO, cells can be predicted in the next frame. Unfortunately, tracker may drift. Therefore we employ a tracker-detector interplay scheme. Each detection around the predicted position is evaluated by calculating the IoU between the bounding box associated with the detection and that associated with the position given by the tracker. For those with an IoU above a threshold (e.g., 50%), cells are tagged as tracked, otherwise lost. For any target cell marked as lost, data association is conducted.

As shown in Fig. 6, the data association module is constructed as a SiameseNet-like network, where the similarity between the target in the previous frame and a detection in current frame is evaluated. As many cells look like similar, using appearance features may not be sufficient to differentiate them, therefore both appearance and motion features are utilized. Two patch pairs are fed into it: a pair of gray scale patches plus a pair of optical flow patches. The similarity between each pair is evaluated, respectively. For each pair, the similarity is compared at three different depths. At each depth, the similarity is evaluated with a structure as shown in Fig. 7.

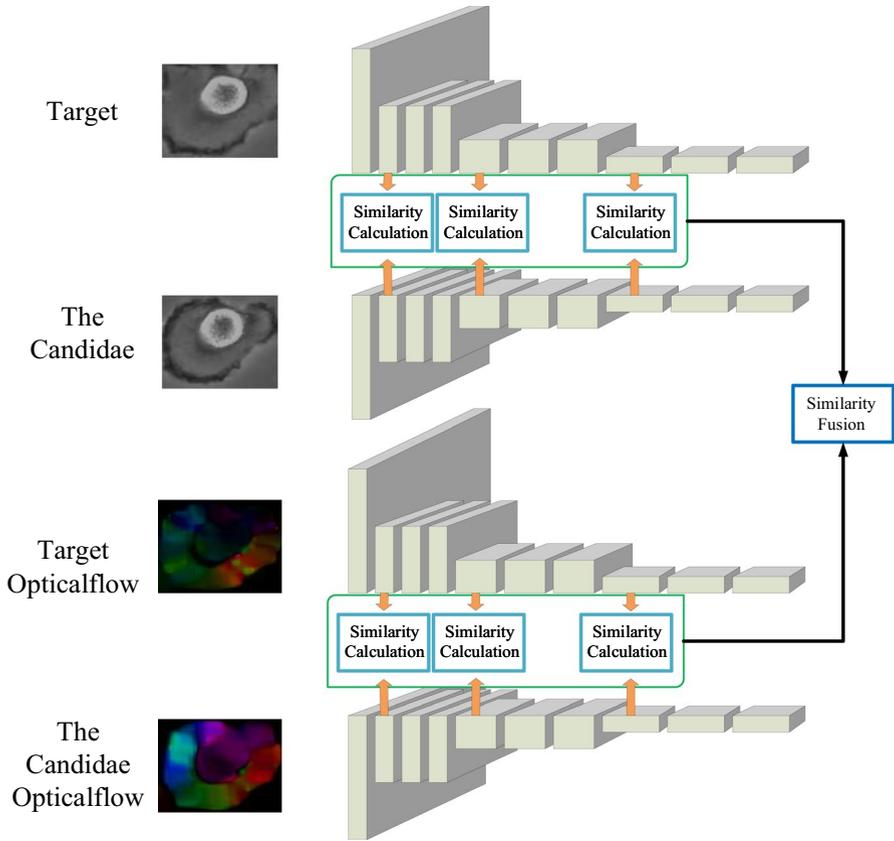


Fig. 6 Data association

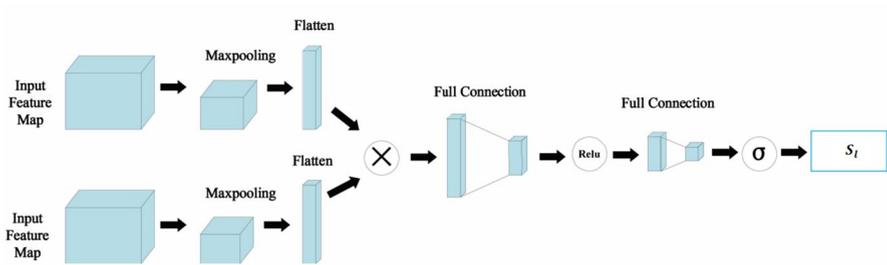


Fig. 7 Similarity calculation module

The similarity score at depth l is denoted as S_l . Finally, in total six different S_l s are combined with their respective weights. Here a uniform weight $\frac{1}{6}$ is used.

$$S_{\text{all}} = \sum_{l=1}^6 \beta_l S_l \quad (11)$$

The data association network is trained with a cross entropy loss:

$$L = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p(y_i)) + (1 - y_i) \log(1 - p(y_i))] \quad (12)$$

where y_i denotes the tag for sample i , $y_i = 1$ denotes that sample i is the same cell, otherwise $y_i = 0$. $p(y_i)$ denotes the predicted tag by the network, and N denotes the batch size. Suppose that there are N target cells in the i th frame, $x^{i,1}, x^{i,2}, x^{i,3}, \dots, x^{i,N}$; and there are M cells detected in the $(i + 1)$ th frame, $x^{i+1,1}, x^{i+1,2}, x^{i+1,3}, \dots, x^{i+1,M}$. For any specific target n in the i th frame, the candidate with the highest $S_{\text{all}}(x^{i,n}, x^{i+1,m})$ in the $(i + 1)$ th frame is thought as the one associated with n .

Those targets in the i th frame, which are successfully associated are marked as “tracked,” and others marked as “lost.” If any target marked as “lost” in the next three continuous frames, i.e., a cell has moved out of the field of view, then its trajectory is terminated. For any new detection, which is not associated with any targets in the i th frame, a new trajectories is initiated.

4 Experimental results

4.1 Experiment setup

All codes are written in Python 3.5.6, TensorFlow-gpu 1.14.0 and Keras 2.2.2. All algorithms are evaluated under Linux Ubuntu 14.2. The hardware platform includes an Intel Xeon Gold 5118 CPU, 256 GB memory, and a Tesla V100 GPU with 16 GB memory.

As many other peer works in cell tracking, our algorithms are evaluated on CTC data sets. Among them, some typical data sets are selected. In the data set Fluo-N2DH-SIM+, cells are hard to detect and segment, as the contrast between the cell and background is not high, and cells look quite similar. In the data set PhC-C2DH-U373(U373), the cells' shape varies drastically, and some impurities are mixed. In the data set Fluo-N2DH-GWOT1(GWOT1), cells are stained unevenly, and cells collide with each other, coming in or out of field of view frequently. In the Fluo-C2DL-Huh7 data set (Huh7), cells are densely populated and their shapes are irregular, and they contact each other.

According to the requirements by CTC (Cell Tracking Challenge) [18] all codes as well as input/output scripts are submitted to CTC for evaluation. CTC will test all submissions with data sets, and then announce the results.

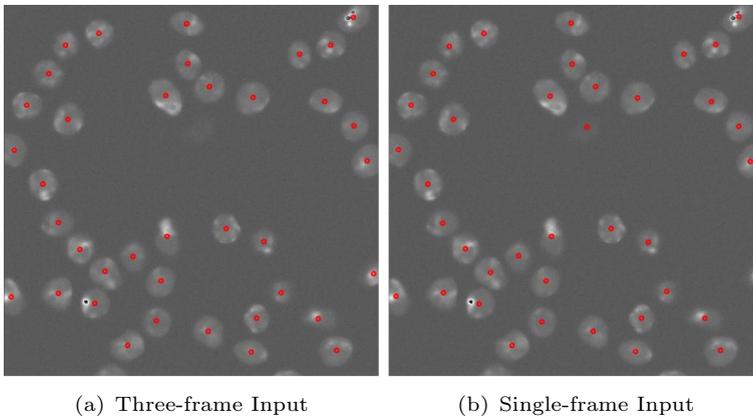


Fig. 8 Effectiveness of attention Unet++

Table 1 AttentionUNet++
Outperforms Unet++

Dataset	Method	DET (Detection)(%)
1	DESU-US [2]	85.5
2	DREX-US [16]	92.8
3	HIT-CN [29]	93.9
4	MON-AU [6]	93.6
5	HD-Wol-GE [22]	87.4
6	OURS	94.5

4.2 Performance on cell detection

The AttentionUnet++ was trained with three neighboring frames, with a backbone ResNet50 [12] is transferred from pretrained ImageNet. The Adam optimizer [14] is used with a learning ratio of 0.0001.

The AttentionUnet++ is trained and tested with the data set SIM+. SIM+-train has 01 and 02 series, and series 01 contains 65 images with a resolution of 628×690, while series 02 contains 150 images with a resolution of 739×773. Both series have a 16-bit gray scale. All these images are normalized and adjusted to 512×512. The SIM+ data set is augmented to triple of its original size. 80% percent of data are used for training, while 20% are used for validation purpose. The output of AttentionUnet++ is centroids of cells, their boundaries are then delineated using a Voronoi algorithm [29]. The performance is scaled with the metric DET.

The effectiveness of our detection algorithms is shown in Fig. 8. As shown in (a), a false detection occurs when a single frame image used, but this does not occur in (b), where multiframe are used. The robustness of multiple frame input is proved.

Table 2 Ablation tests on augmentation and the number of frames

Networks	Aug.	Frames	Precision (%)	Recall (%)	F1 (%)
<i>SIM+ 01 sequence</i>					
Unet++	Yes	3	97.75	98.82	98.28
Att.Unet++	No	3	91.20	98.97	94.93
Att.Unet++	Yes	1	90.91	99.52	95.02
Att.Unet++	Yes	3	97.38	99.40	98.38
<i>SIM+ 02 sequence</i>					
Unet++	Yes	3	83.66	94.63	88.81
Att.Unet++	No	3	45.69	93.04	61.28
Att.Unet++	Yes	1	71.91	95.99	82.22
Att.Unet++	Yes	3	87.08	95.39	91.04

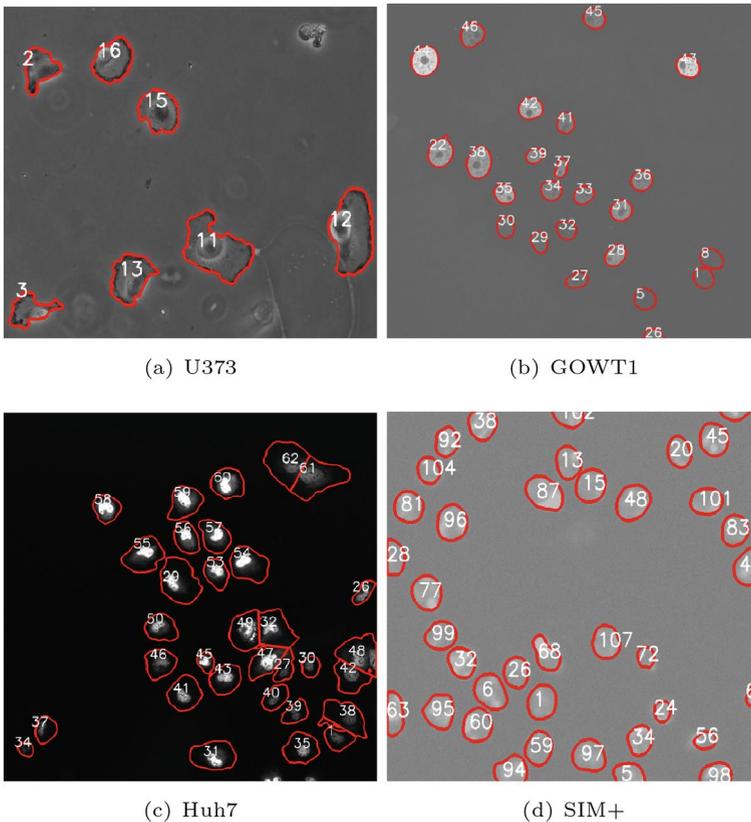


Fig. 9 Cell tracking and segmentation

For the purpose of comparison, some of the famous detection algorithm are listed in Table 1. DESU-US [2] utilizes a temporal-spatial motion diffusion-based Partial Differential Equation (PDE) formulation to detect cells. DREX-US [16]

Table 3 Tracking performance comparison

Dataset	Method	SEG(%)	TRA(%)	OP _{CTB} (%)
Fluo-N2DH-SIM+	DESU-US [2]	64.0	83.7	73.9
	DREX-US [16]	73.1	92.2	82.7
	UP-PT [24]	60.5	89.6	75.0
	OURS	79.7	93.5	86.6
PhC-C2DH-U373	DESU-US	62.5	50.8	56.6
	DREX-US	67.9	93.3	80.6
	IMCB-SG [20]	26.7	95.5	61.1
	UP-PT	35.5	88.3	61.9
Fluo-N2DH-GOWT1	DESU-US	82.3	82.8	82.6
	DREX-US	86.4	91.3	88.9
	IMCB-SG	52.9	88.2	70.5
	UP-PT	72.2	87.5	79.9
Fluo-C2DL-Huh7	DESU-US	85.6	88.8	87.2
	DREX-US	41.5	76.2	58.8
	IMCB-SG	54.0	85.5	69.7
	UP-PT	–	–	–
	OURS	63.3	88.0	75.6

Table 4 Ablation test on ECO and data association

Dataset	ECO	Association	SEG (%)	TRA (%)	OP _{CTB} (%)
U373	No	No	89.74	94.76	92.25
	No	No	90.90	96.57	93.73
	Yes	Yes	90.90	97.13	94.02
GOWT1	No	No	90.61	96.35	93.48
	Yes	No	92.36	97.81	95.09
	Yes	Yes	92.20	98.32	95.26
Huh7	No	No	72.03	94.14	83.09
	Yes	No	73.47	95.38	84.43
	Yes	Yes	75.78	97.44	86.61

uses Gaussian Mixture Model and Euclid Distance Transformation. HIT-CN [29] uses a classic Unet for detection, while MON-AU [6] uses a Mask R-CNN with a backbone of ResNet. HD-Wol-GE [22] extracts features on multiple scales, and then detects cells with a recurrent network. For Table 1, our algorithm outperforms all other algorithms, and the effectiveness of AttentionUnet++ is proven.

Table 5 Ablation test on appearance and motion features

Dataset	Appearance	Motion (%)	TRA(%)
U373	Yes	No	97.04
	No	Yes	96.94
	Yes	Yes	97.13
GOWT1	Yes	No	98.24
	No	Yes	98.02
	Yes	Yes	98.32

As shown in Table 2, ablation tests are conducted on SIM+01 and SIM+02, and metrics of precision, recall and F1 are measured. On SIM+01, the attention mechanism does not improve much. However, on SIM+02, it improves 3.42% on precision, 0.75% on recall and 1.17% on F1. Multiframe input enables the network to fully obtain spatial-temporal information, and the F1 metric is improved about 6.09% on average on SIM+01, and the precision metric is improved by approximately 10.82% on SIM+02. As there are not enough images in both data sets, data augmentation can effectively prevent overfitting. From row 2 and row 4, it can be seen obviously that data augmentation improves all metrics.

4.3 Performance on cell segmentation and tracking

In terms of tracking, the feature compression augment of ECO is set as 14 and 64, σ is set as 0.125, the learning ratio is set as 0.02, and N_s is set as 5. The backbone is transferred from ResNet50, and an Adam optimizer with a learning rate of 0.0001 is utilized.

In Cell Tracking Challenge, tracking performance is scaled with three metrics: SEG, TRA and OP_{CTB} (a combination of SEG and TRA) [18, 19]. SEG denotes the accuracy of segmentation of cells, TRA denotes the overall tracking performance, and OP_{CTB} is the average of SEG and TRA. The tracking performance of our proposed method is compared with some state-of-the-art methods, such as traditional methods or deep learning-based methods. The tracking and segmentation effectiveness can be shown in Fig. 9. Segmentation is performed using the method proposed in [29], which can delineate the cell's boundary very well.

As shown in Table 3, due to our powerful cell detection algorithm and multimodal feature fusion, our proposed ECOMA algorithm outperforms all others except DREX-US, which is slightly better than our algorithm on GOWT1. In summary, our algorithm is more robust.

The contributions of ECO and our Data Association are evaluated as shown in Table 4. From the table, it can be found that the first row corresponds to the Nearest Neighbor (NN) algorithm, which is used as the base line. Compared with the greedy NN, it can be seen that multimode feature fusion improves 1.14% on SEG, 1.5% on TRA, and 1.47% on OP_{CTB} . Data association improves 1.04% on TRA, 0.81% on OP_{CTB} . It seems that the data association strategy works but not very much. One possible reason is that cells does not move very far from frame

to frame, the IoUs are sufficient for tracking. If the frame rate is reduced, and the cell's position in nearby frames does not overlap, then the advantage of tracking and data association will become obvious.

From Table 5, it can be shown that the combination of appearance and motion features outperforms any single feature, but appearance features contribute more to the tracking performance, motion features can be helpful to enhance the tracking performance, but not much.

5 Conclusion

In this research, we proposed AttentionUnet++ detector as well as ECOMA tracker. The AttentionUnet++ detector takes in multi-frame input, and utilizes spatial and temporal features. Extending single-target tracker ECO to multi-cell tracking, the proposed ECOMA tracker utilizes both appearance and motion features and a Sames-like data association module is included. Experiments are conducted on CTC data set, the efficiency of our algorithms is verified, indicating that our algorithms can outperform many highly ranked algorithms published by CTC. Future works may focus on a more robust data association module, which can evaluate multiple frames and can deal with broken tracklets or tracklets connected by mistake.

Acknowledgements The authors wish to thank Natural Science Foundation of Guangdong for their support through Grant Number 2020A1515010706.

Author Contributions FW designed the overall algorithm framework, and wrote the main manuscript text. HL and WY wrote python code for Sects. 2 and 3, respectively. SJ prepared most figures and data collection. PG conducted proofreadings.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethical approval Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Amat F, Lemon W, Mossing DP et al (2014) Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data. *Nat Methods* 11(9):951–958
2. Arbelles A, Reyes J, Chen JY et al (2018) A probabilistic approach to joint cell tracking and segmentation in high-throughput microscopy videos. *Med Image Anal* 47:140–152
3. Boquet-Pujadas A, Olivo-Marin JC, Guillén N (2021) Bioimage analysis and cell motility. *Patterns* 2(1):100–170
4. Bouguet JY et al (2001) Pyramidal implementation of the affine Lucas Kanade feature tracker description of the algorithm. *Intel Corp* 5(1–10):4
5. Boukari F, Makrogiannis S (2018) Automated cell tracking using motion prediction-based matching and event handling. *IEEE/ACM Trans Comput Biol Bioinform* 17(3):959–971
6. Chang MJ (2021) Mon-au. <http://celltrackingchallenge.net/participants/mon-au/> Accessed 2021
7. Dai C, Zhang Z, Huang J et al (2018) Automated non-invasive measurement of single sperm's motility and morphology. *IEEE Trans Med Imaging* 37(10):2257–2265
8. Danelljan M, Robinson A, Shahbaz Khan F, et al (2016) Beyond correlation filters: Learning continuous convolution operators for visual tracking. In: *European Conference on Computer Vision*, Springer, pp 472–488
9. Danelljan M, Bhat G, Shahbaz Khan F et al (2017) Eco: efficient convolution operators for tracking. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 6638–6646
10. Gao P, Ma Y, Song K et al (2018) High performance visual tracking with circular and structural operators. *Knowl-Based Syst* 161:240–253
11. Giusti A, Cireşan DC, Masci J, et al. (2013) Fast image scanning with deep max-pooling convolutional neural networks. In: *2013 IEEE International Conference on Image Processing*, IEEE, pp 4034–4038
12. He K, Zhang X, Ren S, et al. (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 770–778
13. He T, Mao H, Guo J et al (2017) Cell tracking using deep neural networks with multi-task learning. *Image Vis Comput* 60:142–153
14. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980)
15. Kwak YH, Hong SM, Park SS (2010) A single cell tracking system in real-time. *Cellular Immunol* 265(1):44–49
16. Layton Aho AC, Raymond Y (2021) <http://celltrackingchallenge.net/participants/DREX-US/> Accessed 2021
17. Li R, Gao Q, Rohr K (2021) Multi-object dynamic memory network for cell tracking in time-lapse microscopy images. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, IEEE, pp 1029–1032
18. Maška M, Ulman V, Svoboda D et al (2014) A benchmark for comparison of cell tracking algorithms. *Bioinformatics* 30(11):1609–1617
19. Matula P, Maška M, Sorokin DV et al (2015) Cell tracking accuracy measurement based on comparison of acyclic oriented graphs. *PLoS One* 10(12):e0144959
20. Pengdong Xiao WY (2012) Imcb-sg (2). [https://public.celltrackingchallenge.net/participants/IMCB-SG%20\(2\).pdf](https://public.celltrackingchallenge.net/participants/IMCB-SG%20(2).pdf) Accessed 2021
21. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, pp 234–241
22. Royden Wagner KR (2021) Hd-wag-ge. <http://celltrackingchallenge.net/participants/HD-Wag-GE/> Accessed 2021
23. Sandler M, Howard A, Zhu M, et al (2018) Mobilenetv2: inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 4510–4520
24. Tiago E, Pedro Q, Maja T-O (2015) Up-pt, segmentation & tracking. <http://celltrackingchallenge.net/participants/UP-PT/> Accessed 2015
25. Valiūškaitė V, Raudonis V, Maskeliūnas R et al (2020) Deep learning based evaluation of spermatozoid motility for artificial insemination. *Sensors* 21(1):72

26. Winter MR, Fang C, Banker G et al (2012) Axonal transport analysis using multitemporal association tracking. *Int J Comput Biol Drug Design* 5(1):35–48
27. Yue Y, Zong W, Li X et al (2020) Long-term, in toto live imaging of cardiomyocyte behaviour during mouse ventricle chamber formation at single-cell resolution. *Nat Cell Biol* 22(3):332–340
28. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, et al (2018) Unet++: a nested u-net architecture for medical image segmentation. In: *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, pp 3–11
29. Zhou Z, Wang F, Xi W et al (2019) Joint multi-frame detection and segmentation for multi-cell tracking. In: *International Conference on Image and Graphics*, Springer, pp 435–446

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.