Computers in Biology and Medicine 146 (2022) 105543

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

TOD-CNN: An effective convolutional neural network for tiny object detection in sperm videos

Shuojia Zou [a], Chen Li [a,*], Hongzan Sun [b], Peng Xu [c], Jiawei Zhang [a], Pingli Ma [a], Yudong Yao [d], Xinyu Huang [e], Marcin Grzegorzek [e]

[a] Microscopic Image and Medical Image Analysis Group, College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China
[b] Shengjing Hospital, China Medical University, Shenyang, China
[c] Jinghua Hospital, Shenyang, China
[d] Department of Electrical and Computer Engineering, Stevens Institute of Technology, USA
[e] Institute of Medical Informatics, University of Luebeck, Luebeck, Germany

ARTICLE INFO

Keywords:
Image analysis
Object detection
Convolutional neural network
Sperm microscopy video

ABSTRACT

The detection of tiny objects in microscopic videos is a problematic point, especially in large-scale experiments. For tiny objects (such as sperms) in microscopic videos, current detection methods face challenges in fuzzy, irregular, and precise positioning of objects. In contrast, we present a convolutional neural network for tiny object detection (TOD-CNN) with an underlying data set of high-quality sperm microscopic videos (111 videos, > 278,000 annotated objects), and a graphical user interface (GUI) is designed to employ and test the proposed model effectively. TOD-CNN is highly accurate, achieving 85.60% AP_{50} in the task of real-time sperm detection in microscopic videos. To demonstrate the importance of sperm detection technology in sperm quality analysis, we carry out relevant sperm quality evaluation metrics and compare them with the diagnosis results from medical doctors.

1. Introduction

Sperm is necessary for the human and mammal reproductive process, which plays an important role in human reproduction and animal breeding [1]. With the continuous development of computer technology, researchers have tried to use computer-aided image analysis in many fields, such as whole-slide image analysis [2], histopathology image analysis [3–5], cytopathological analysis [6,7], COVID-19 image analysis [8,9], and microorganism counting [10]. In addition, in the field of semen analysis and diagnosis, researchers have also proposed many *Computer Aided Semen Analysis* (CASA) systems [11]. As the first step of the CASA system, sperm detection is one of the most important parts to support the reliability of sperm analysis results [12]. At present, most sperm detection techniques [13–16] are based on traditional image processing techniques such as thresholding, edge detection and contour fitting. However, for many techniques, the detection results require manual intervention. The common difficulties for sperm detection mainly include the small size, uncertain morphologies and low contrast of the sperms, which are difficult for locating. Moreover, there are lots of similar impurities in the samples for misleading (as shown in Section 4 Fig. 7).

In recent years, more and more excellent object detection models are constantly proposed [17], such as *Region-based CNN* (*R*-CNN) series models [18–21], *You Only Look Once* (YOLO) series models [22–25], *Single Shot Multibox Detector* (SSD) [26], and RetinaNet [27]. The performance of *Convolutional Neural Networks* (CNN) has obviously surpassed the complex classic image processing algorithms in the field of medical image processing [28–30], which makes it possible to use deep learning methods to perform real-time sperm object detection tasks in sperm microscopic videos. However, the accuracy of sperm object detection is still lower than that of object detection under conventional scales [31]. Hence, techniques such as feature fusion and residual networks are used in our method to improve the detection performance in this field. The technologies above are applied to build an easy-to-operate sperm detection model (TOD-CNN), and an AP_{50} of 85.60% is achieved in the task of sperm detection for microscopic videos.

The workflow of the proposed TOD-CNN detection method is summarized as follows (as shown in Section 3 Fig. 1): (a) Training and Validation Data: The training and validation data contains 80 sperm microscopic videos and corresponding annotation data with the location

* Corresponding author.
E-mail address: lichen@bmie.neu.edu.cn (C. Li).

https://doi.org/10.1016/j.compbiomed.2022.105543
Received 1 March 2022; Received in revised form 30 March 2022; Accepted 17 April 2022
Available online 22 April 2022
0010-4825/© 2022 Elsevier Ltd. All rights reserved.

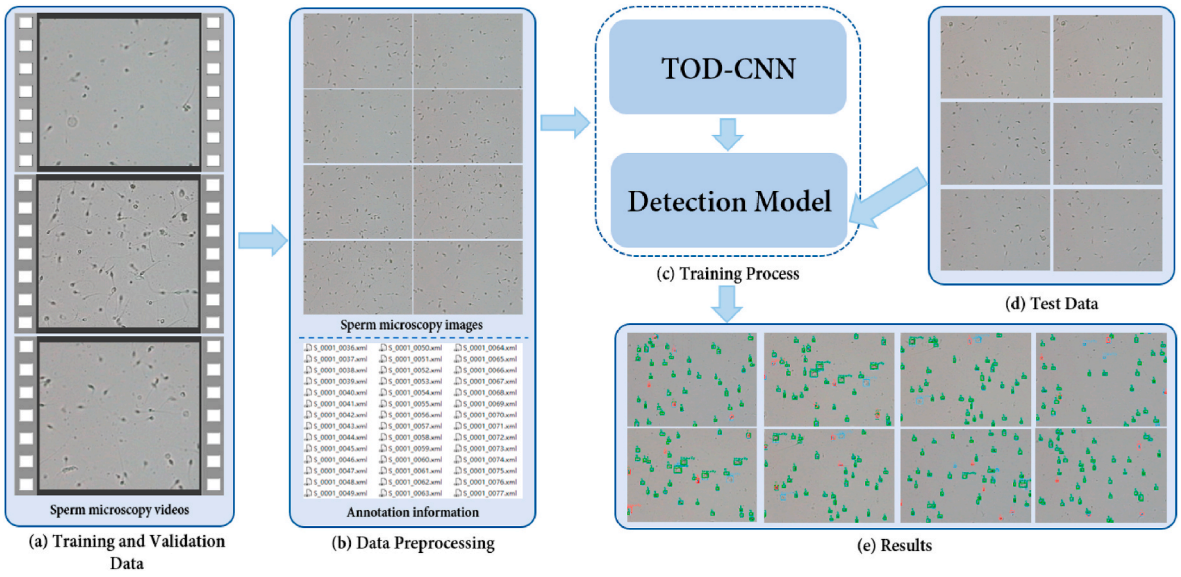

Fig. 1. The workflow of the proposed sperm object detection method using TOD-CNN.

and category information of sperms and impurities. (b) Data Preprocessing: The sperm microscopic video is divided into frames to obtain one by one sperm microscopic images, and the object information is annotated by using LabelImg software. (c) Training Process: The TOD-CNN model is trained and the best model is saved to perform sperm object detection. (d) Test Data: The test data contains 21 sperm microscopic videos.

The main contributions of this paper are as follows:

- Build an easy-to-operate CNN for sperm detection, namely TOD-CNN (Convolutional Neural Network for tiny object detection).
- TOD-CNN has excellent detection results and real-time detection ability in the task of tiny object detection in sperm microscopic video, achieving 85.60% AP_{50} and 35.7 *frames per second* (FPS).

The structure of this paper is as follows: Section 2 introduces the existing sperm object detection methods based on traditional methods, machine learning methods, and deep learning methods. Section 3 illustrates the detailed design of TOD-CNN. Section 4 introduces the data set used in the experiment, experiment settings, evaluation methods, and results. Chapter 5 is conclusion.

2. Related work

2.1. Existing sperm object detection methods

2.1.1. Traditional methods

Traditional methods mainly include three types, which are threshold-based methods, shape fitting methods, and filtering methods. Threshold-based methods: Urbano et al. [14] use Gaussian filter to enhance the image, and then the image is binarized using the Otsu [32] threshold method, and the result is morphologically operated to determine the position of the sperm; Elseyed et al. [13] use several certain frames to generate the background information, then the background information is subtracted from the original image (to suppress noise). Finally, the Otsu threshold is applied to determine the position of the sperm. Shape fitting methods: Zhou et al. [33] use a rectangular area which is similar to the shape of the object (sperm) to fit the object, and then the position of the sperm is described by the parameters of the rectangle. Yang et al. [16] use an ellipse to approximate the sperm head, and the improved multiple birth and cut algorithm based on marked point processes [34] is used to detect and locate the head of the sperm through modelling. Filtering methods: Ravanfar et al. [35] select several suitable structural elements firstly, and then the operation based on Top-hat is used to filter the image sequence to achieve the purpose of separating sperm and other debris. Nurhadiyatna et al. [36] use the *Gaussian Mixture Model* (GMM) enhanced by the Hole Filling Algorithm as the probability density function to predict the probability of each pixel in the image belongs to the foreground and the background. The researchers found that the calculation amount of this method is significantly less than other methods.

2.1.2. Machine learning methods

The unsupervised learning method is the most used machine learning method. Berezansky et al. [37] use the Spatio-Temporal Segmentation to detect sperm by segmentation, integrating *k*-means, GMM, mean shift, and other segmentation methods. Shi et al. [38] use the optical capture method for sperm detection.

2.2. Deep learning based object detection methods

Deep learning methods are widely used in many artificial intelligent fields, for example classification [39–42], segmentation [43–45] and object detection [46,47]. Furthermore, some widely recognized general object detection models that have been proposed in recent years are introduced bellow.

2.2.1. One-stage object detector

The YOLO series models [22–25], RetinaNet [27], and SSD [26] are prominent representatives of one-stage object detectors. The one-stage object detectors are based on the idea of regression, which can directly output the final prediction results from the input images without generating suggested regions in advance. YOLO series models: They use Darknet as the backbone of the model to extract features from the image. The v1, v2, v3, v4 are successively proposed by improving the backbone network structure, improving the loss function, using batch normalization, feature pyramid network [48], spatial pyramid pooling network [49], and other optimization methods. RetinaNet: It uses ResNet [50] and the feature pyramid network as the backbone of the model, whose main contribution is proposing a focus loss function. The focus loss function solves the imbalance between the number of foreground and background categories in a single-stage object detector. SSD: It uses VGG16 [51] as the basic model and then adds a new convolutional layer based on VGG16 to obtain more feature maps for detection and

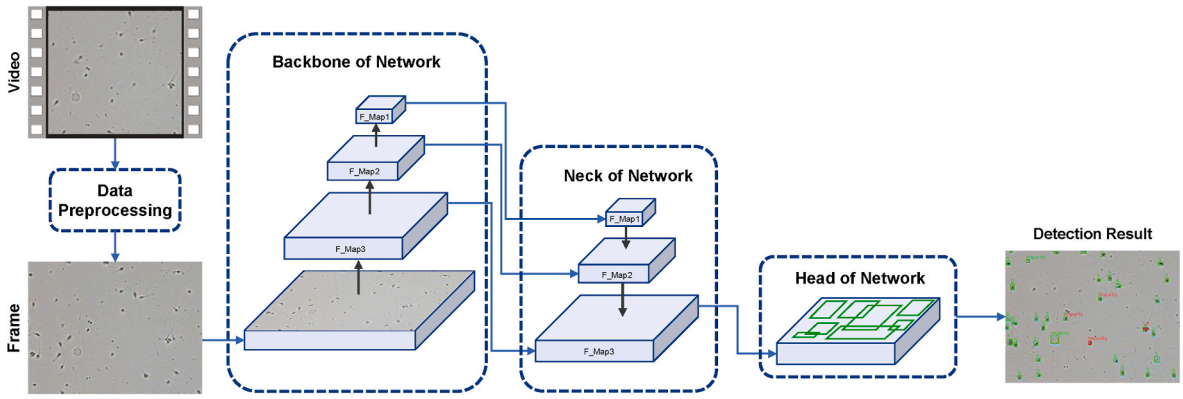

Fig. 2. The architecture of TOD-CNN.

generates the final prediction result by fusing the prediction results of 6 feature maps.

2.2.2. *Two-stage object detector*

The models based on *R*–CNN series are classical two-stage object detectors. The *R*–CNN [18] generates proposal regions through the selective search algorithm. Then the features of the proposal regions are extracted by using CNN. Finally, the SVM classifier is used to predict the objects in each region and identify the category of the objects. The Fast *R*–CNN [19] no longer extracts features for each proposal region. The features of entire image is extracted using CNN, then each proposal region and corresponding features are mapped. Besides, the Fast *R*–CNN uses a multi-task loss function, allowing us to train the detector and bounding box regressor simultaneously. The Faster *R*–CNN [20] replaces the selective search algorithm with Region Proposal Network, which can help CNN to generate proposal regions and detect objects simultaneously.

3. TOD-CNN based sperm detection method in microscopic image

Sperm detection is always the first step in a CASA system, which determines the reliability of the results of sperm microscopic video analysis. However, the existing algorithms cannot accurately detect sperms. Therefore, we follow the idea of YOLO [22–25], ResNet [50], Inception-v3 [52], and VGG16 [51] models and propose a novel one-stage deep learning based sperm object detection model (TOD-CNN). The workflow of the proposed TOD-CNN detection approach is shown in Fig. 1.

3.1. *Basic knowledge*

In this section, the methods related to our work are introduced, including YOLO, ResNet, Inception-v3, and VGG16 models.

3.1.1. *Basic knowledge of YOLO*

YOLO series models solve the object detection task as a regression problem. The YOLO series models remove the step of generating the proposal region in the two-stage object detector and accelerate the detection process. YOLO-v3 [24] is the most popular model in the YOLO series models due to its excellent detection performance and speed.

The YOLO-v3 model mainly consists of four parts, which are preprocessing, backbone, neck, and head. The preprocessing: The *k*-means algorithm is used to cluster nine anchor boxes in the data set before training. The backbone: YOLO-v3 uses Darknet53 as the backbone network of the model. Darknet53 does not have maxpooling layers and fully connected layers. The fully convolutional network can change the size of the tensor by changing the strides of the convolutional kernel. In addition, Darknet53 follows the idea of ResNet and adds a residual module to the network to solve the vanishing gradient problem of the deep network. The neck: The neck of the YOLO-v3 model draws on the idea of the feature pyramid network [48] to enrich the information of the feature map. The head: The head of YOLO-v3 outputs 3 feature maps with different sizes and then detects large, medium, and small size objects of the three feature maps with three sizes.

3.1.2. *Basic knowledge of ResNet*

ResNet [50] is one of the most widely used feature extraction CNNs due to its practical and straightforward structure. With the continuous deepening of CNN, the model's performance cannot be continuously improved, and the accuracy may even decrease. However, ResNet proposes the Shortcut Connection structure to solve the problems above. The identity mapping operation and residual mapping operation are included in the Shortcut Connection structure. The identity mapping is to pass the current feature map backward through cross-layer transfer (when the dimension of feature map does not match, a 1×1 convolution operation is used to adjust the dimension of feature map). Residual mapping is to pass the current feature map to the next layer after convolution operation. A Shortcut Connection structure contains one identity mapping operation and two or three residual mapping operations in general.

3.1.3. *Basic knowledge of Inception-v3 and VGG16*

In Inception-v3 [52], to reduce the parameters and ensure the performance of the model, an operation that replaces $N \times N$ convolution kernels with $1 \times N$ and $N \times 1$ convolution kernels is proposed. The receptive fields of $1 \times N$ and $N \times 1$ convolution kernels and $N \times N$ convolution kernels are the same, where the former has less parameters than the latter. In addition, the Inception-v3 model can support multi-scale input, which can use convolution kernels with different sizes to perform convolution operations on the input images, and then the input feature maps can be connected to generate the final feature map.

VGG16 [51] model includes 13 convolutional layers, 3 fully connected layers, and 5 maxpooling layers. The most prominent feature of the VGG16 model is its simple structure. All convolutional layers use the same convolution kernel parameters, and all pooling layers use the same pooling kernel parameters. Although the VGG16 model has a simple structure, it has strong feature extraction capabilities.

3.2. *The structure of TOD-CNN*

The TOD-CNN model, which refers to the YOLO series model, can regard the object detection task as a regression problem for fast and precise detection. The architecture of TOD-CNN is shown in Fig. 2, where the entire network is composed of four parts: Data preprocessing, backbone of the network, neck of the network and head of the network.

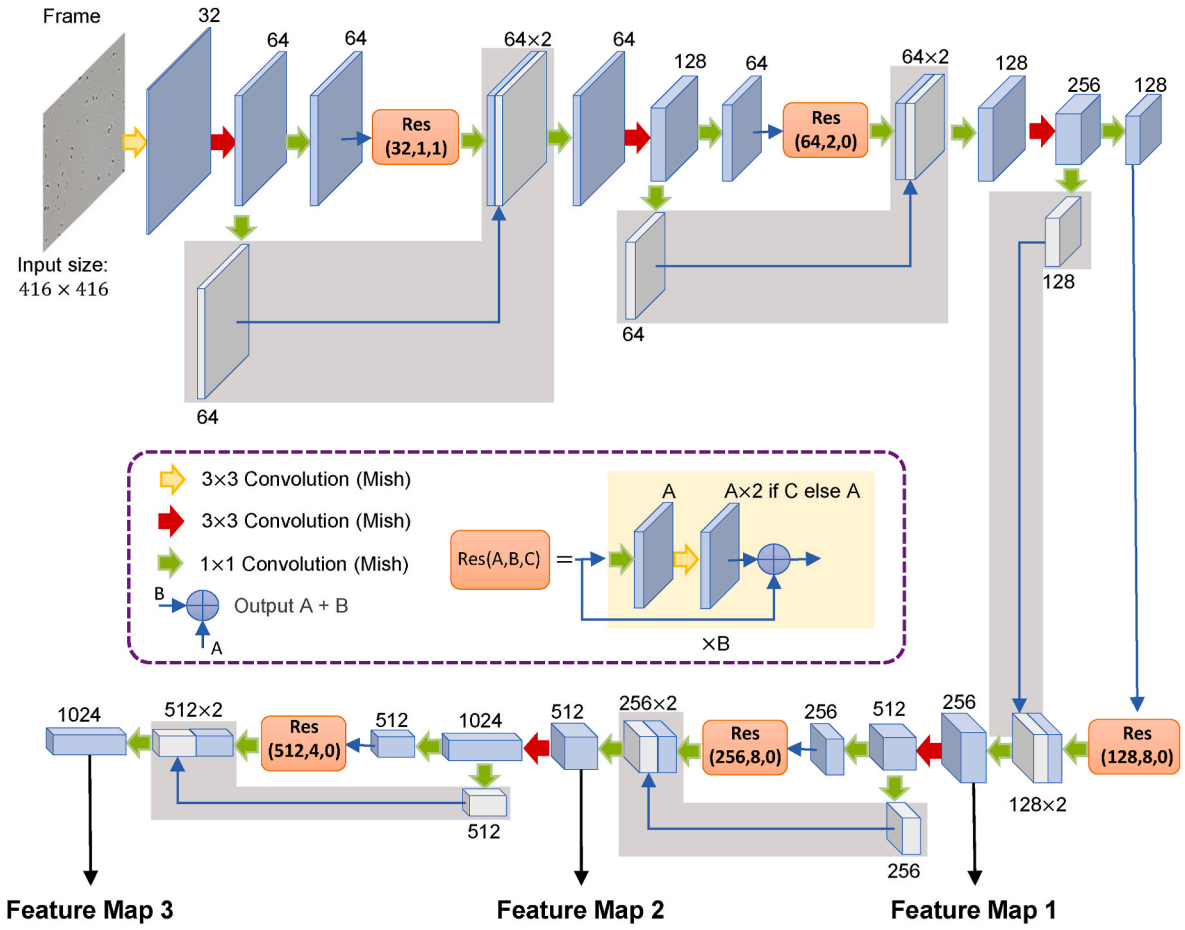

Fig. 3. The architecture of TOD-CNN backbone.

The detailed implementation of each part is introduced in detail below.

3.2.1. *Data Preprocessing*

The object detection task in the video is essentially based on image processing. Therefore, it is necessary to split the sperm microscopic video into continuous frames (single images). However, due to the movement of the lens during the sperm microscopic video shooting process, there are some blurred frames in the sperm microscopic video. After analysing the grayscale histogram of frames, there is an obvious difference between the grayscale distribution of the blurred frame and the normal frame. Therefore, the blurred frames can be solved by deleting images whose Otsu threshold is less than a certain threshold (from artificial experience). In addition, TOD-CNN is an anchor-based object detection model. Therefore, the *k*-means algorithm is used to cluster a certain number (TOD-CNN uses six) of anchor boxes in data set to train the model.

3.2.2. *The backbone of TOD-CNN*

A straight forward backbone structure with cross-layer concatenate operation is designed, which is shown in Fig. 3. However, in a fully convolutional network, as the structures of CNNs continue to deepen, the semantic information of the feature map becomes more and more

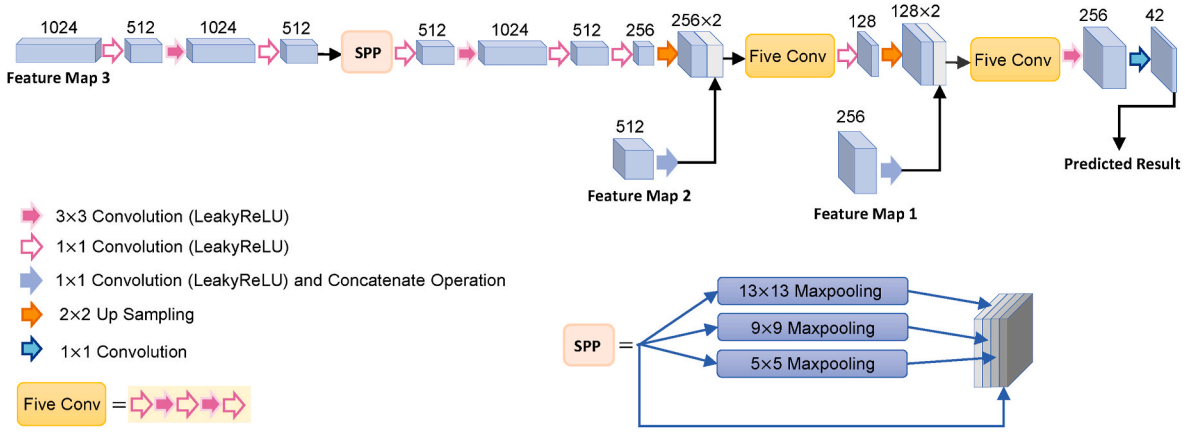

Fig. 4. The architecture of TOD-CNN neck.

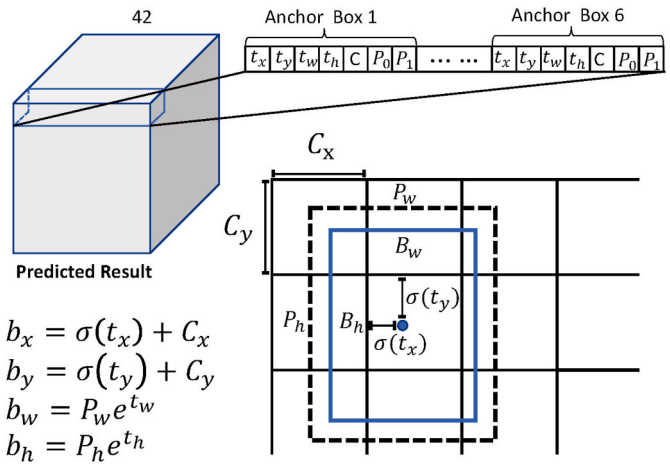

Fig. 5. The architecture of TOD-CNN head. Calculate the coordinates of the prediction box using the network output result and priori boxes. t_x, t_y, t_w and t_h are predicted by TOD-CNN for locating the bounding box. P_0 and P_1 represent the probability of sperm and impurity in the bounding box, respectively. C is the confidence to determine whether there is an object in the bounding box.

abundant, while the location information of the feature map constantly decreases. As a result, the network can improve the classification performance but may reduce the positioning accuracy. Our work focuses on detecting tiny objects and accurate locating, which needs to maintain precise local information. Therefore, we enhance the transfer of location information (transferring shallow features to deep layers) through the following methods: First, we refer to the residual idea of ResNet, the Shortcut Connection structure provides the approach for transferring local information with a cross-layer add operation, which is used in TOD-CNN (as shown in Res (A, B, C) in Fig. 3); second, based on the straightforward backbone structure, a cross-layer concatenate operation is applied to enhance the transfer of local information (as shown in grey shaded part in Fig. 3).

The detailed design of TOD-CNN backbone is shown in Fig. 3, where the input size of the backbone is 416 × 416, the yellow arrow indicates convolution operations with a kernel size of 3 × 3 and stride of 1 (each use filtering with padding and followed by a Mish activation), the red arrow indicates convolution operations with a kernel size of 3 × 3 and stride of 2 (each followed by a Mish activation), and the green arrow indicates convolution operations with a kernel size of 1 × 1 and stride of 1 (each use filtering with padding and followed by a Mish activation).

3.2.3. The neck of TOD-CNN

In object detection model, the main purpose of model neck is to integrate feature informations extracted from model backbone. The neck structure of TOD-CNN is shown in Fig. 4. In fact, there are abnormal morphological sperms (very big) and some other impurities (such as bacteria, protein lumps and bubbles) in semen. These sperms and impurities are significantly different from normal sperm in size. Therefore, to collect multi-scale information, we have adopted spatial pyramid pooling operation [49] to integrate multi-scale information into TOD-CNN neck. In addition, due to the small sizes of tiny objects, the information of tiny objects might be easily lost in down-sampling process. In order to solve this problem, the feature fusion method is used in TOD-CNN neck, where the shallow and deep feature maps are fused by upsampling to avoid the loss of tiny object information.

The detailed design of TOD-CNN neck are shown in Fig. 4, where all convolution operations are with stride of 1 (each use filtering with padding), and the detailed illustration of the kernel size and activation function is shown in Fig. 4. Finally, TOD-CNN neck outputs a feature map of size (input size/8) × (input size/8) × 42, where 42 is the number of anchor boxes (6) × 7, because each anchor box needs to have 7 parameters: the relative center coordinates, the width and the high offset, class, and confidence, the details are explained in Section 3.2.4.

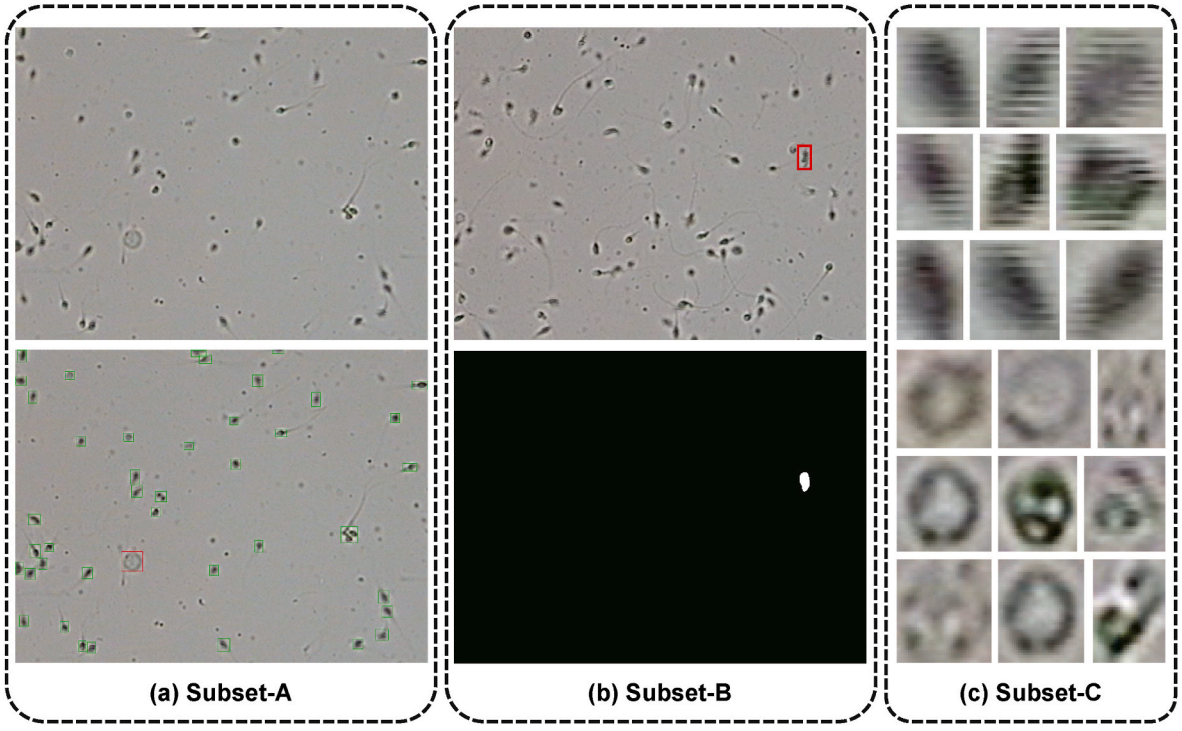

Fig. 6. An example of the sperm video data set. (a) The first row shows a frame in a sperm microscopic video and the bottom row is the corresponding annotation for object detection tasks. Sperms are in green boxes and impurities are in red boxes. (b) The first row shows a frame in sperm microscopic video and the bottom row shows the corresponding ground truth for object tracking tasks. (c) The first row shows individual sperm images and the bottom row shows individual impurity images for classification tasks.

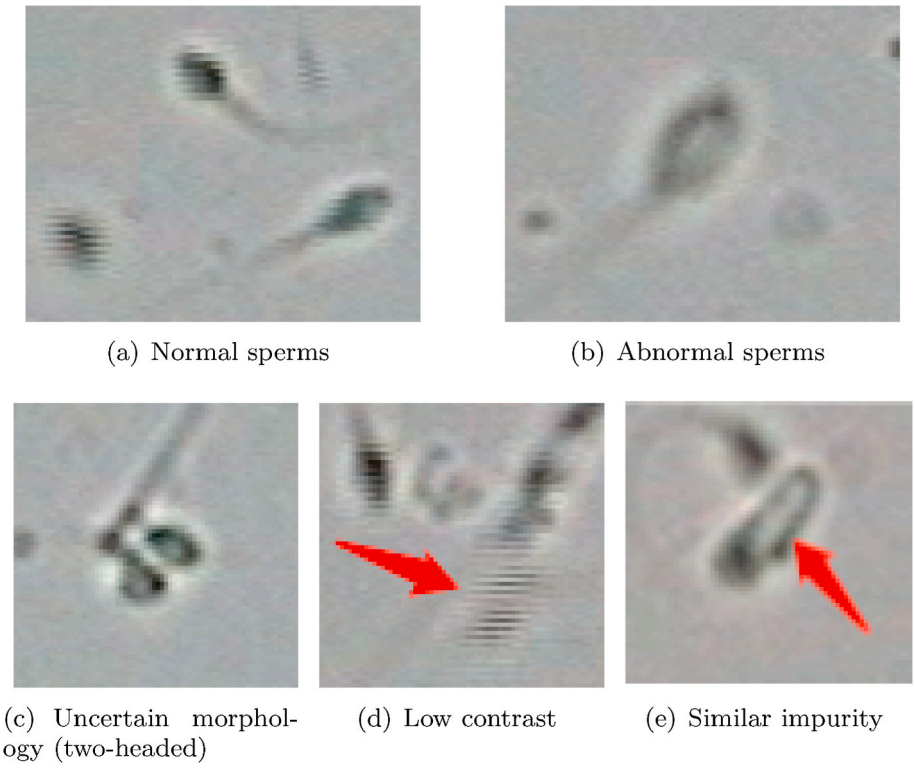

Fig. 7. Challenging cases for the detection of sperms. The positions pointed by the red arrow are the blur of sperm imaging caused by sperm movement and the impurity similar to sperm.

3.2.4. The head of TOD-CNN

In the head of TOD-CNN, 6 bounding boxes are predicted for each cell in the output feature map. For each bounding box, 7 coordinates (t_x, t_y, t_w, t_h, C, P_0 and P_1) are predicted, so the dimension of the predicted result in Fig. 5 is 42. For each cell, the offset from the upper left corner of the image is assumed to be (C_x, C_y), and the width and height of the corresponding a priori box are P_w and P_h. The calculation method of center coordinates (b_x and b_y), width (b_w) and height (b_h) of predicted box is shown in Fig. 5. Multi-label classification is applied to predict the categories in each bounding box. Furthermore, due to the dense prediction method is applied to the head of TOD-CNN, the non-maximum value suppression method based on distance intersection over union [53] is used to remove bounding boxes with high overlap in the output results of the network.

4. Experiments

4.1. Experimental settings

4.1.1. Data set

A sperm microscopic video data set is released in our previous work [54] and it is used for the experiments of this paper. These sperm microscopic videos in the data set are obtained by a WLJY-9000 computer-aided sperm analysis system [55] under a 20 × objective lens and a 20 × electronic eyepiece. More than 278,000 objects are annotated in the data set: normal, needle-shaped, amorphous, cone-shaped, round, or multi-nucleated head sperms and impurities (such as bacteria, protein clumps, and bubbles). The object sizes range from approximately 5 to 50 μm^2. These objects are annotated by 14 reproductive doctors and biomedical scientists and verified by 6 reproductive doctors and biomedical scientists.

From 2017 to 2020, the collection and preparation of this data set took four years, including more than 278,000 annotated objects, as shown in Fig. 6. Furthermore, the data set contains some hard-to-detect objects, such as uncertain morphology sperm, low contrast sperm, and similar impurities (as show in Fig. 7), which greatly increases the difficulty of tiny object detection.

In this data set, Subset-A provides more than 125,000 objects with bounding box annotation and category information in 101 videos for tiny object detection task; Subset-B segments more than 26,000 sperms in 10 videos as ground truth for tiny object tracking task; Subset-C provides more than 125,000 independent images of sperms and impurities for tiny object classification task. Although Subset-C is not used in this work, it is still openly available to non-commercial scientific work.

4.1.2. Training, validation, and test data setting

We randomly divide the sperm microscopic video into training, validation, and test data sets at a ratio of 6:2:2. Therefore, we have 80 sperm microscopic videos and corresponding annotation information for training, and validation. The training set includes 2125 sperm microscopic images (77522 sperms and 2759 impurities), and validation set includes 668 sperm microscopic images (23173 sperms and 490 impurities). And we have 21 sperm microscopic videos for testing, the test set includes 829 sperm microscopic images (20706 sperms and 1230 impurities).

4.1.3. Experimental environment

The experiment is conducted by Python 3.7.0 in Windows 10 operating system. The models we use in this paper are implemented by Keras 2.1.5 framework with Tensorflow 1.13.1 as the backend. Our experiment uses a workstation with Intel$^{(R)}$ Core$^{(TM)}$ i7-9700 CPU with 3.00 GHz, 32 GB RAM, and NVIDIA GEFORCE RTX 2080 8 GB.

4.1.4. Hyper parameters

The purpose of object detection task is to find all objects of interest in the image. Therefore, this task can be regarded as a combination of positioning and classification tasks. Therefore, as the loss function of the network, we use the complete intersection over union [53] (CIoU)

Table 1
The definitions of evaluation metrics, where TP, TN, FP and FN represent True Positive, True Negative, False Positive and False Negative, respectively; N denotes the number of detected objects.

Metric	Definition	Metric	Definition
Average Precision	$\frac{\sum_{i=1}^{N}\{Precision(i)\times Recall(i)\}}{Number\ of\ Annotations}$	Recall	$\frac{TP}{TP+FN}$
F_1 Score	$2\times\frac{Precision\times Recall}{Precision+Recall}$	Precision	$\frac{TP}{TP+FP}$

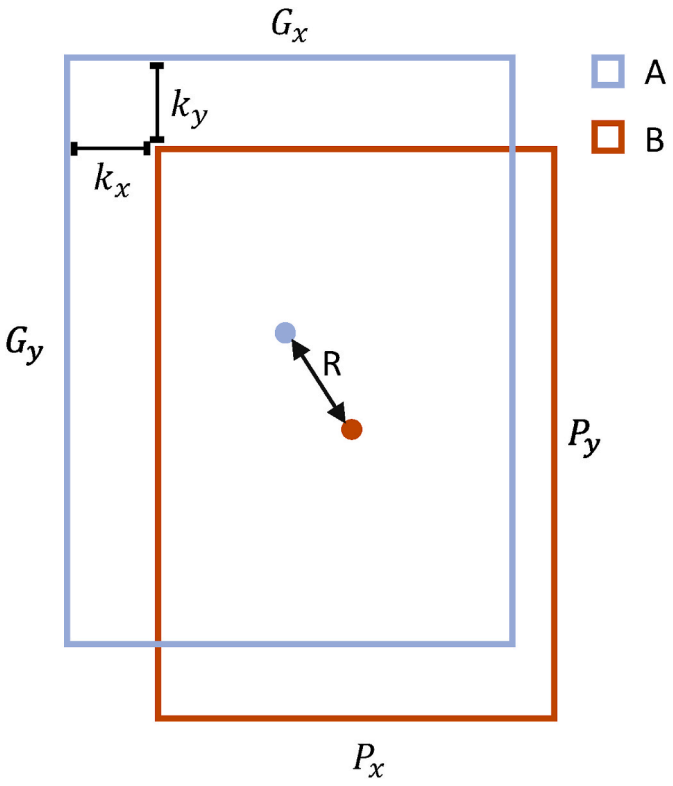

Fig. 8. The IoU calculation method.

function (location loss function) and the binary cross-entropy function (confidence and classification loss function), and then minimize them by Adam optimizer. For other hyper parameters, when freezing part of the layer training and unfreezing all layers, the batch size is set to 16 and 4, the training is 50 and 100 epochs, and the learning rate is set to 1×10^{-3} and 1×10^{-4}, respectively. Besides, the cosine annealing scheduler [56] is used to adjust the learning rate. Besides, when the loss value no longer drops, the training is terminated early.

4.2. Evaluation metrics

In order to quantitatively compare the performance of various object detection methods, different metrics are used to evaluate the detection results. Recall (Rec), Precision (Pre), F_1 Score (F_1), and Average Precision (AP) which can be used to evaluate the detection results.

Rec measures how many objects present in the annotation information are correctly detected. However, we cannot judge the detection result from the perspective of Rec alone. Pre measures how many objects detected by the model exist in the annotation information. The F_1 is the harmonic average of model Pre and Rec, and is an metric used to measure model performance. AP is a metric, which is widely used to evaluate the performance of object detection models. It can be obtained by calculating the area under the curve of Pre and Rec. It can evaluate object detection models from two aspects: Pre and Rec. The definitions of these evaluation metrics are provided in Table 1.

Table 2
The memory costs, training time and FPS of TOD-CNN, YOLO-v4, YOLO-v3, SSD, RetinaNet and Faster *R*–CNN.

Model	Memory Cost	Training Time	FPS
TOD-CNN	164 MB	**119 min**	35.7
YOLO-v4	244 MB	135 min	28.4
YOLO-v3	235 MB	374 min	**37.0**
SSD	**91.2 MB**	280 min	31.5
RetinaNet	139 MB	503 min	21.0
Faster *R*–CNN	108 MB	2753 min	7.8

The metrics in Table 1 are calculated based on True Positive, True Negative, False Positive, and False Negative. The intersection over union (IoU) is one of the evaluation criteria for evaluating whether the detected object is positive or negative. The calculation method of IoU is shown in Fig. 8 and Eq. (1).

$$IoU = \frac{A\bigcap B}{A\bigcup B} = \frac{(G_x-k_x)(G_y-k_y)}{G_xG_y+P_xP_y-(G_x-k_x)(G_y-k_y)} \tag{1}$$

From Fig. 8 and Eq. (1), it can be found that the smaller the values of G_x, G_y, P_x, and P_y, the more sensitive the value of IoU to the changes of k_x and k_y. The above phenomenon further illustrates that it is very difficult to detect tiny objects and it is unfair to use IoU alone to evaluate tiny object detection. Therefore, without affecting the sperm positioning, we propose a more suitable evaluation index. This indicator is a positive sample when the detected object meets two conditions at the same time: the first is that the detected object category is correct, and the second is that the IoU of the detection box and the ground truth box exceeds $B1$, or the IoU of the detection box and the ground truth box exceeds $B2$, and the distance between the center points of the two box does not exceed R pixels.

4.3. Evaluation of sperm detection methods

In order to prove the effectiveness of the proposed TOD-CNN method for sperm detection in sperm microscopic videos, we compared its detection results with other state-of-the-art methods, such as YOLO-v3 [24], YOLO-v4 [25], SSD [26], RetinaNet [27], and Faster *R*–CNN [20]. In the experiment process, each metric is calculated under the condition of $B1=0.5$, $B2=0.45$, and $R=3$. $B1$ and $B2$ represent the IoU value, and R represents the pixel distance between the center of the predicted box and the center of the ground-truth box. Among them, whether $B1$ is greater than 0.5 is a more common standard for evaluating positive and negative samples in the field of object detection [57]. In addition, after our extensive experimental verification, the sperm object center coordinates obtained when $B2\geq0.45$ and $R\leq3$ have little effect on the sperm tracking task. Therefore, this paper adopts this standard to evaluate the experimental results.

4.3.1. Compare with other methods

In this part, we make a comparison between TOD-CNN and some state-of-the-art methods in terms of memory costs, training time, FPS, and detection performance.

4.3.1.1. Evaluation of memory, time costs and FPS. To compare the memory costs, training time and FPS among TOD-CNN, YOLO-v4, YOLO-v3, SSD, RetinaNet and Faster *R*–CNN, we provide the details in Table 2.

From Table 2, we can find that the memory cost of TOD-CNN is 164 MB, the training time of TOD-CNN is around 119 min for 60 sperm microscopy videos, and the FPS is 34.7. In contrast, the memory cost and FPS of TOD-CNN are not optimal, but it considers both the model size

Table 3
A comparison of detection results between TOD-CNN and existing models. (In [%].)

Models	AP	F1	Pre	Rec
TOD-CNN	**85.60**	**90.00**	89.47	**90.54**
YOLO-v4	51.00	70.16	85.19	59.64
YOLO-v3	42.93	64.36	78.36	54.60
SSD	65.00	78.51	93.48	67.67
RetinaNet	15.05	27.00	**95.62**	15.72
Faster RCNN	35.76	55.28	46.57	67.99

and real-time performance. By comparing with YOLO-v3 and YOLO-v4, TOD-CNN has the minor memory cost. By comparing with RetinaNet and Faster *R*–CNN, TOD-CNN has faster detection speed. By comparing with SSD, TOD-CNN does not have better memory cost and real-time performance, but sperm detection ability of TOD-CNN is much better than SSD, which will be explained in detail in the next paragraph.

4.3.1.2. Evaluation of sperm detection performance. TOD-CNN is compared with existing object detection models using our data set. In Table 3, we list the comparison with the best performance results of various models (YOLO-v4, YOLO-v3, SSD, RetinaNet, and Faster *R*–CNN). In TOD-CNN, AP is nearly 20% higher, F1 is nearly 12% higher and Rec is nearly 22% higher. Our Pre is about 6% lower than the best performing model (RetinaNet). It is observed that our Rec is 75% higher than that of RetinaNet, which shows that the number of detected objects obtained by TOD-CNN far exceeds RetinaNet. Overall, TOD-CNN outperforms existing models in sperm detection. Furthermore, a visual comparison of the models discussed above is shown in Fig. 9.

From Fig. 9, we can see that the correct detection case of TOD-CNN is only fewer than Faster RCNN in "sperm-lack" scenes (oligospermia), but our Pre is much higher than Faster *R*CNN [58]. In "sperm-normal" scenes (healthy), the correct detection case of TOD-CNN is the best and our Pre and Rec are higher. By observing Fig. 9, it is easy to understand why TOD-CNN has slightly lower Pre than SSD and RetinaNet, but other metrics are better than other models (the number of correct detections of

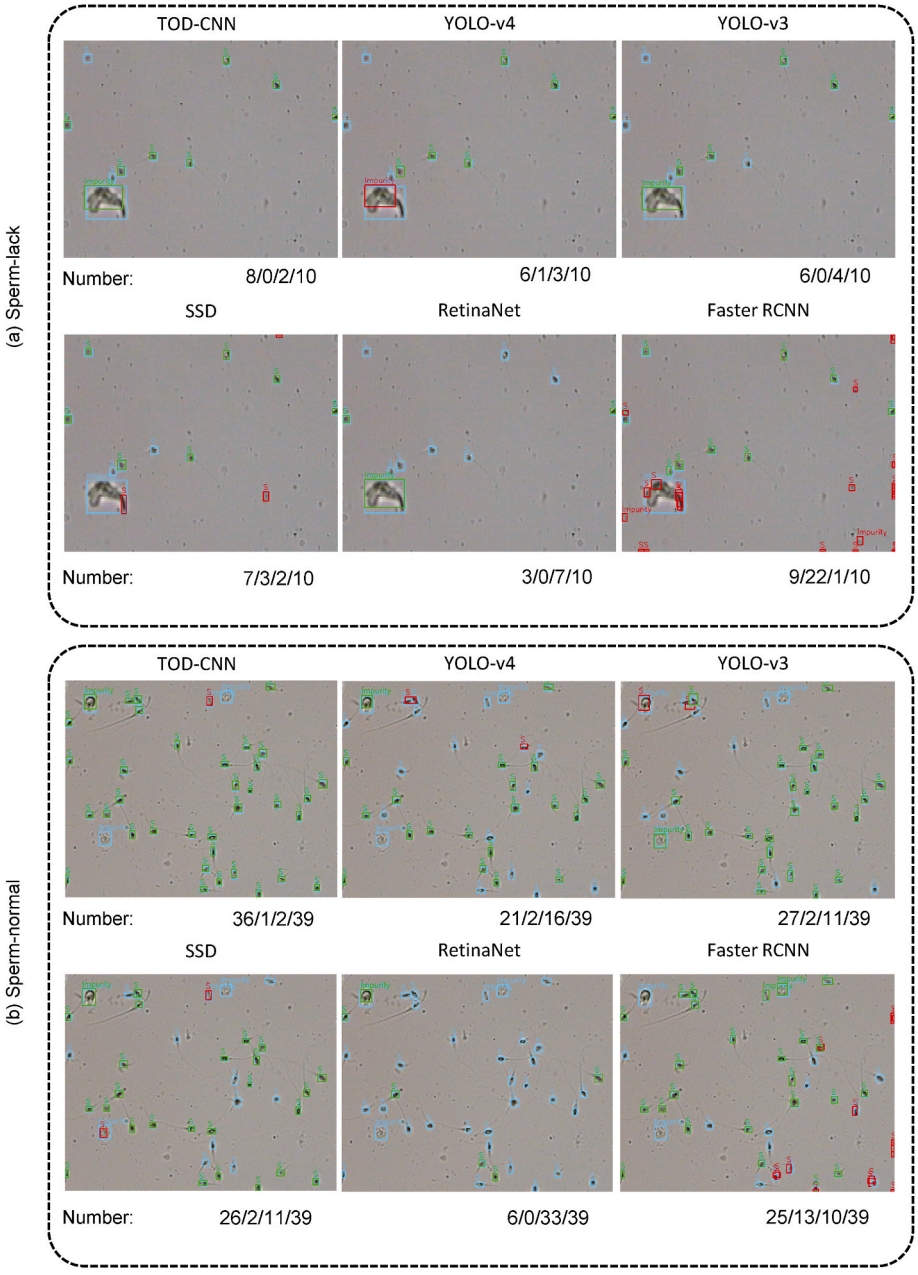

Fig. 9. Comparison of TOD-CNN with YOLO-v4 [25], YOLO-v3 [24], SSD [26], RetinaNet [27], and Faster RCNN [58]. In these images, the blue boxes represent the corresponding ground-truth, the green boxes correspond to the correctly detected objects, and the red boxes correspond to the incorrectly detected objects. The values represent the number of correctly detected objects/the number of incorrectly detected objects/the number of objects in the annotation information but are not detected/and the total number of objects in the annotation information.

Table 4
A comparison between TOD-CNN and existing models in the scene with impurity, where AP_S, AP_I and mAP represents AP of sperm, AP of impurity and mean AP, respectively. (In [%].)

Models	AP_S	AP_I	mAP	F1	Pre	Rec
TOD-CNN	**85.60**	**57.33**	**71.47**	**88.57**	88.41	**88.74**
YOLO-v4	51.00	30.00	40.50	69.61	84.76	59.06
YOLO-v3	42.93	35.90	39.42	63.80	78.34	53.81
Faster RCNN	35.76	25.80	30.78	54.52	46.06	66.78
SSD	65.00	18.95	41.98	76.59	92.23	65.44
RetinaNet	15.05	33.84	24.44	28.51	**95.36**	16.76

Table 5
The detection results, μ and STDEV of the five-fold cross-validation experiments. (In [%].)

Metrics	AP	F1	Pre	Rec
1	85.60	90.00	89.47	90.54
2	84.90	90.11	92.76	87.61
3	88.80	92.78	95.19	90.48
4	86.37	91.33	94.12	88.66
5	87.29	90.65	91.15	90.16
μ	86.59	90.97	92.54	89.49
STDEV	1.36	1.02	2.05	1.16

TOD-CNN far exceeds other models).

Furthermore, to test the robustness of TOD-CNN against impurities in the microscopic videos, we have added 4479 impurities into the experiments. The experimental results are shown in Table 4, where TOD-CNN shows the best robustness against the effect of impurities compared to other models.

4.3.2. Cross-validation experiment

To verify the reliability, stability and repeatability of TOD-CNN, we have performed five-fold cross-validation. The experimental results are shown in Table 5, where the mean values (μ) of the four evaluation metrics is higher than 89% except for AP, and AP is higher than 86%. It can be seen that TOD-CNN has good performance and repeatability. The standard deviation (STDEV) of F1 is 1.02%, the STDEV of two of the four evaluation metrics are below 1.40%, and only the STDEV of Pre is slightly higher (2.05%), showing that TOD-CNN is relatively stable and reliable.

4.3.3. Sperm tracking

The ultimate goal of sperm detection is to find the sperm trajectories and calculate the relevant parameters for clinical diagnosis. TOD-CNN and two models with better detection results (YOLO-v4 and SSD) are compared for sperm tracking in Table 3. Based on the detection result of each model, we use the *k*NN algorithm to match sperms in adjacent video frames to the actual trajectories marked in Subset-B. The visualization results are shown in Fig. 10. We can observe that our tracking trajectories are very close to the actual ones, and the trajectory discontinuity or incorrect tracking is rarely occurred due to the stronger detection capability of TOD-CNN.

In addition, we calculate three important motility parameters of sperms on Subset-B, including the Straight Line Velocity (VSL), Curvilinear Velocity (VCL) and Average Path Velocity (VAP) [59] of actual trajectories, with TOD-CNN, SSD and Yolo-v4, respectively. Comparing with the actual trajectories, the error rates of VSL, VCL and VAP calculated with TOD-CNN (10.15%, 5.09% and 8.95%) are significantly lower than that of SSD (41.58%, 5.01% and 17.40%) and Yolo-v4 (12.73%, 36.12% and 19.65%). Based on VCL, VSL, and VAP, an experienced threshold value from a clinical doctor is set to determine whether a sperm is motile to calculate the corresponding progressive motility (PR). The error between PR obtained by TOD-CNN tracking results and doctors' diagnosis results are all within 9%. The experimental result shows that our TOD-CNN can assist doctors in clinical work.

4.3.4. A python-based graphical user interface

To conveniently use TOD-CNN to detect tiny objects in microscopic videos and images, we design a Python-based GUI (Fig. 11) that can help users to control the Intersection of Union (IoU) threshold and confidence according to their own needs to achieve the desired test performance. Besides, users can load Model Path to use their own setting/weights for tiny object detection. This GUI is compatible with videos (such as ".mp4" and ".avi") and images (such as ".png" and ".jpg") in various formats.

5. Conclusion and discussion

We develop and present a public, massive and high-quality data set for sperm detection, tracking and classification, and this data set now is published and available online. We also provide a one-stage CNN model (TOD-CNN) on Subset-A for tiny object detection in real-time, which can accurately detect sperms in videos and images. However, TOD-CNN fails in some cases and cannot detect sperms completely or accurately. The

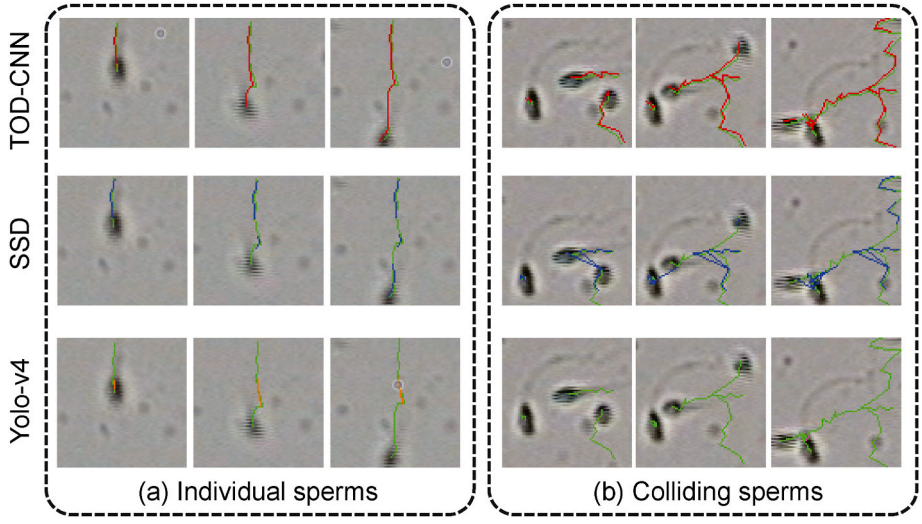

Fig. 10. An example of sperm tracking results. Green lines represent the actual trajectories based ground truth; red, blue and orange lines denote the tracking trajectories of TOD-CNN, SSD and Yolo-v4, respectively.

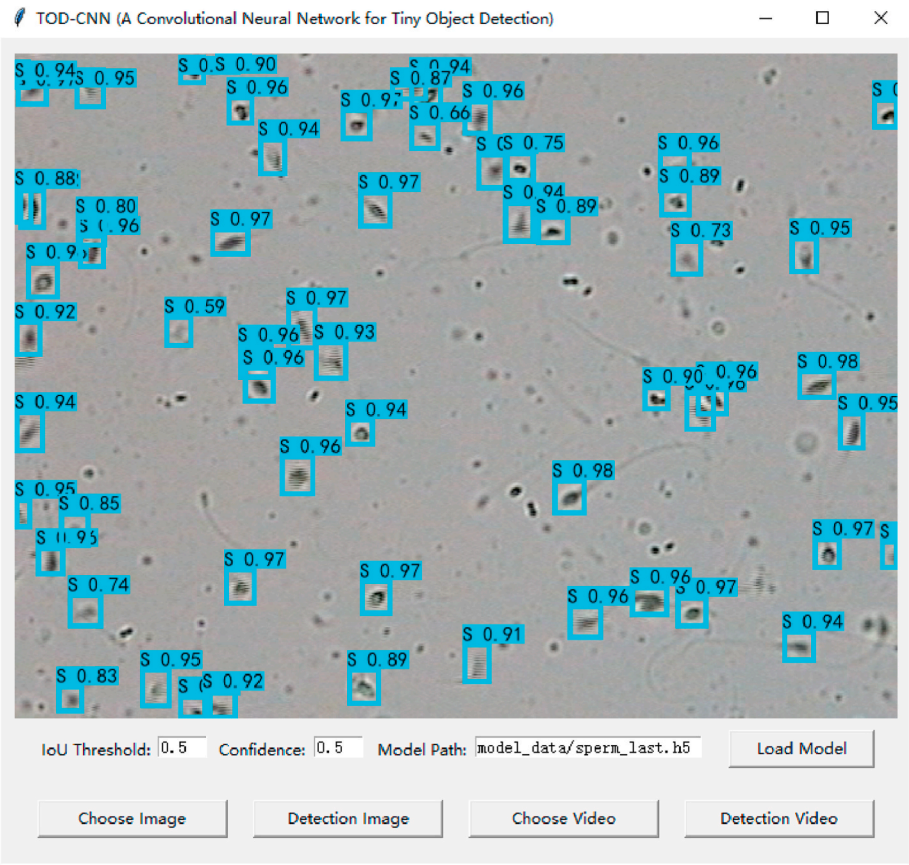

Fig. 11. GUI of TOD-CNN for detecting sperms in microscopic videos or images.

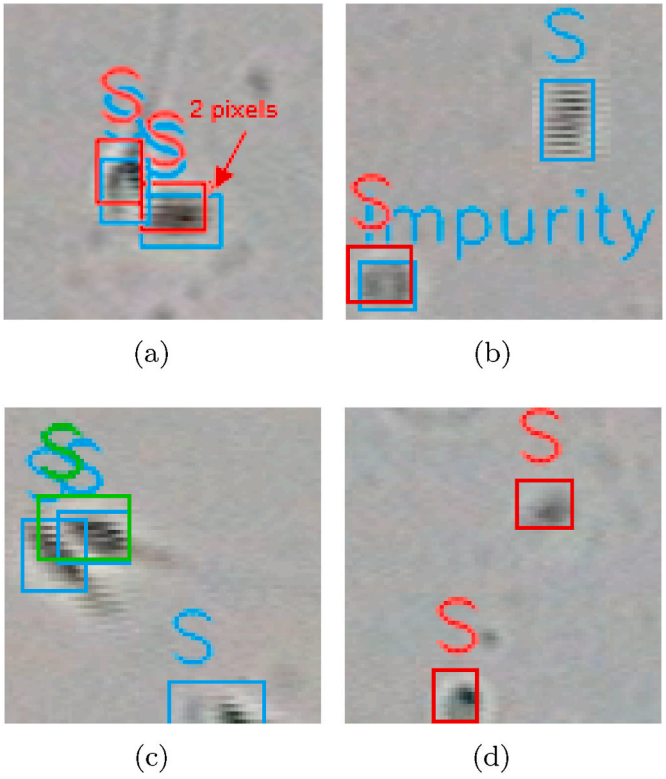

Fig. 12. Visualized results of some typical detection failures of TOD-CNN. The red and green boxes represent the detection results, the blue boxes represent the ground truth, S represents sperms and Impurity represents impurities.

example of incorrect detection results are shown in Fig. 12.

In Fig. 12(a), we can see that the detection boxes can surround sperms correctly. However, due to the small size of the ground truth boxes, a minor position offset (one or two pixels) causes the IoU between detection and ground truth boxes to be lower than 0.5. In Fig. 12(b), due to the movement of the sperms, the thickness of the semen wet film and noticeable interference fringes in sperm videos, it may lose valuable information and lead to errors in detection. Also, because some impurities have very close visual information to sperms, TOD-CNN incorrectly detects the impurities as sperms. In Fig. 12(c), for the sperms appearing on figure edges, it is difficult to explore the complete information and sometimes these sperms are missed in detection. To ensure the annotation information reliability, when we marked sperms in videos, we only choose sperms without controversy. In Fig. 12(d), the detected sperms may be located deeply in the semen wet film. Because of its unclear imaging, it is difficult to distinguish whether it is a sperm or an impurity and it is not annotated in our data set.

In future work, we will continue to integrate related optimization algorithms to improve the performance of TOD-CNN, such as monarch butterfly optimization [60], earthworm optimization algorithm [61], elephant herding optimization [62,63], moth search algorithm [64], slime mould algorithm [65], hunger games search [66], Runge Kutta optimizer [67], colony predation algorithm [68,69], and Harris hawks optimization [70].

Declaration of competing interest

The authors declare that they have no conflict of interest.

Acknowledgements

This work is supported by the “National Natural Science Foundation

of China" (No. 61806047). We thank Miss Zixian Li and Mr. Guoxian Li for their important discussion.

References

[1] S. Gadadhar, G. Alvarez Viar, J.N. Hansen, A. Gong, A. Kostarev, C. Ialy-Radio, S. Leboucher, M. Whitfield, A. Ziyyat, A. Touré, Tubulin glycylation controls axonemal dynein activity, flagellar beat, and male fertility, Science 371 (6525) (2021), eabd4914, https://doi.org/10.1126/science.abd4914.

[2] X. Li, C. Li, M.M. Rahaman, H. Sun, X. Li, J. Wu, Y. Yao, M. Grzegorzek, A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches, Artif. Intell. Rev. (2022) 1–70, https://doi.org/10.1007/s10462-021-10121-0.

[3] Y. Li, X. Wu, C. Li, X. Li, H. Chen, C. Sun, M.M. Rahaman, Y. Yao, Y. Zhang, T. Jiang, A hierarchical conditional random field-based attention mechanism approach for gastric histopathology image classification, Appl. Intell. (2022) 1–22, https://doi.org/10.1007/s10489-021-02886-2.

[4] Y. Li, C. Li, X. Li, K. Wang, M.M. Rahaman, C. Sun, H. Chen, X. Wu, H. Zhang, Q. Wang, A comprehensive review of markov random field and conditional random field approaches in pathology image analysis, Arch. Comput. Methods Eng. 29 (1) (2022) 609–639, https://doi.org/10.1007/s11831-021-09591-w.

[5] C. Li, H. Chen, X. Li, N. Xu, Z. Hu, D. Xue, S. Qi, H. Ma, L. Zhang, H. Sun, A review for cervical histopathology image analysis using machine vision approaches, Artif. Intell. Rev. 53 (7) (2020) 4821–4862, https://doi.org/10.1007/s10462-020-09808-7.

[6] M.M. Rahaman, C. Li, Y. Yao, F. Kulwa, X. Wu, X. Li, Q. Wang, Deepcervix: a deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques, Comput. Biol. Med. 136 (2021) 104649, https://doi.org/10.1016/j.compbiomed.2021.104649.

[7] M.M. Rahaman, C. Li, X. Wu, Y. Yao, Z. Hu, T. Jiang, X. Li, S. Qi, A survey for cervical cytopathology image analysis using deep learning, IEEE Access 8 (2020) 61687–61710, https://doi.org/10.1109/ACCESS.2020.2983186.

[8] M.M. Rahaman, C. Li, Y. Yao, F. Kulwa, M.A. Rahman, Q. Wang, S. Qi, F. Kong, X. Zhu, X. Zhao, Identification of covid-19 samples from chest x-ray images using deep learning: a comparison of transfer learning approaches, J. X Ray Sci. Technol. 28 (5) (2020) 821–839, https://doi.org/10.3233/XST-200715.

[9] C. Li, J. Zhang, F. Kulwa, S. Qi, Z. Qi, A sars-cov-2 microscopic image dataset with ground truth images and visual ffatures, in: Pattern Recognition and Computer Vision, PRCV), 2020, pp. 244–255, https://doi.org/10.1007/978-3-030-60633-6_20.

[10] J. Zhang, C. Li, M.M. Rahaman, Y. Yao, P. Ma, J. Zhang, X. Zhao, T. Jiang, M. Grzegorzek, A comprehensive review of image analysis methods for microorganism counting: from classical image processing to deep learning approaches, Artif. Intell. Rev. (2021) 1–70, https://doi.org/10.1007/s10462-021-10082-4.

[11] W. Zhao, P. Ma, C. Li, X. Bu, S. Zou, T. Jang, M. Grzegorzek, A survey of semen quality evaluation in microscopic videos using computer assisted sperm analysis, arXiv preprint arXiv:2202.07820, https://doi.org/10.48550/arXiv.2202.07820, 2022.

[12] W. Zhao, S. Zou, C. Li, J. Li, J. Zhang, P. Ma, Y. Gu, P. Xu, X. Bu, A survey of sperm detection techniques in microscopic videos, in: The Fourth International Symposium on Image Computing and Digital Medicine, 2020, pp. 219–224, https://doi.org/10.1145/3451421.3451467.

[13] M. Elsayed, T.M. El-Sherry, M. Abdelgawad, Development of computer-assisted sperm analysis plugin for analyzing sperm motion in microfluidic environments using image-j, Theriogenology 84 (8) (2015) 1367–1377, https://doi.org/10.1016/j.theriogenology.2015.07.021.

[14] L.F. Urbano, P. Masson, M. VerMilyea, M. Kam, Automatic tracking and motility analysis of human sperm in time-lapse images, IEEE Trans. Med. Imag. 36 (3) (2017) 792–801, https://doi.org/10.1109/TMI.2016.2630720.

[15] X. Li, C. Li, F. Kulwa, M.M. Rahaman, W. Zhao, X. Wang, D. Xue, Y. Yao, Y. Cheng, J. Li, S. Qi, T. Jiang, Foldover features for dynamic object behaviour description in microscopic videos, IEEE Access 8 (2020) 114519–114540, https://doi.org/10.1109/ACCESS.2020.3003993.

[16] H. Yang, X. Descombes, S. Prigent, G. Malandain, X. Druart, F. Plouraboué, Head tracking and flagellum tracing for sperm motility analysis, in: 2014 IEEE 11th International Symposium on Biomedical Imaging, ISBI), 2014, pp. 310–313, https://doi.org/10.1109/ISBI.2014.6867871.

[17] Z. Zou, Z. Shi, Y. Guo, J. Ye, Object detection in 20 years: a survey, arXiv preprint arXiv:1905.05055, https://doi.org/10.48550/arXiv.1905.05055, 2019.

[18] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR), 2014, pp. 580–587, https://doi.org/10.1109/CVPR.2014.81.

[19] R. Girshick, Fast r-cnn, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440–1448, https://doi.org/10.1109/ICCV.2015.169.

[20] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149, https://doi.org/10.1109/TPAMI.2016.2577031.

[21] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988, https://doi.org/10.1109/ICCV.2017.322.

[22] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR), 2016, pp. 779–788, https://doi.org/10.1109/CVPR.2016.91.

[23] J. Redmon, A. Farhadi, Yolo9000: better, faster, stronger, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR), 2017, pp. 6517–6525, https://doi.org/10.1109/CVPR.2017.690.

[24] J. Redmon, A. Farhadi, Yolov3: an incremental improvement, arXiv preprint arXiv: 1804.02767, https://doi.org/10.48550/arXiv.1804.02767, 2018.

[25] A. Bochkovskiy, C.Y. Wang, H.Y.M. Liao, Yolov4: optimal speed and accuracy of object detection, arXiv preprint arXiv:2004.10934, https://doi.org/10.48550/arXiv.2004.10934, 2020.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, Ssd: single shot multibox detector, in: Computer Vision – ECCV 2016, 2016, pp. 21–37, https://doi.org/10.1007/978-3-319-46448-0_2.

[27] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: 2017 IEEE International Conference on Computer Vision, ICCV), 2017, pp. 2999–3007, https://doi.org/10.1109/ICCV.2017.324.

[28] B. Gu, R. Ge, Y. Chen, L. Luo, G. Coatrieux, Automatic and robust object detection in x-ray baggage inspection using deep convolutional neural networks, IEEE Trans. Ind. Electron. 68 (10) (2021) 10248–10257, https://doi.org/10.1109/TIE.2020.3026285.

[29] L. Wang, M. Shen, C. Shi, Y. Zhou, Y. Chen, J. Pu, H. Chen, Ee-net: an edge-enhanced deep learning network for jointly identifying corneal micro-layers from optical coherence tomography, Biomed. Signal Process Control 71 (2022) 103213, https://doi.org/10.1016/j.bspc.2021.103213.

[30] W. Yang, H. Zhang, J. Yang, J. Wu, X. Yin, Y. Chen, H. Shu, L. Luo, G. Coatrieux, Z. Gui, Q. Feng, Improving low-dose ct image using residual convolutional network, IEEE Access 5 (2017) 24698–24705, https://doi.org/10.1109/ACCESS.2017.2766438.

[31] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, S. Yan, Perceptual generative adversarial networks for small object detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR), 2017, pp. 1951–1959, https://doi.org/10.1109/CVPR.2017.211.

[32] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybernet. 9 (1) (1979) 62–66, https://doi.org/10.1109/TSMC.1979.4310076.

[33] X. Zhou, Y. Lu, Efficient mean shift particle filter for sperm cells tracking, in: 2009 International Conference on Computational Intelligence and Security, 2009, pp. 335–339, https://doi.org/10.1109/CIS.2009.264.

[34] E. Soubiès, P. Weiss, X. Descombes, A 3d segmentation algorithm for ellipsoidal shapes. application to nuclei extraction, in: ICPRAM-international Conference on Pattern Recognition Applications and Methods, 2013, pp. 97–105. https://hal.archives-ouvertes.fr/hal-00733187.

[35] M.R. Ravanfar, M.H. Moradi, Low contrast sperm detection and tracking by watershed algorithm and particle filter, in: 2011 18th Iranian Conference of Biomedical Engineering, ICBME), 2011, pp. 260–263, https://doi.org/10.1109/ICBME.2011.6168568.

[36] A. Nurhadiyatna, A.L. Latifah, D. Fryantoni, T. Wirahman, R. Wijayanti, F. H. Muttaqien, Comparison and implementation of motion detection methods for sperm detection and tracking, in: 2014 International Symposium on Micro-NanoMechatronics and Human Science, MHS), 2014, pp. 1–5, https://doi.org/10.1109/MHS.2014.7006125.

[37] M. Berezansky, H. Greenspan, D. Cohen-Or, O. Eitan, Segmentation and tracking of human sperm cells using spatio-temporal representation and clustering, in: Medical Imaging 2007: Image Processing, 2007, pp. 891–902, https://doi.org/10.1117/12.708887.

[38] L.Z. Shi, J. Nascimento, C. Chandsawangbhuwana, M.W. Berns, E.L. Botvinick, Real-time automated tracking and trapping system for sperm, Microsc. Res. Tech. 69 (11) (2006) 894–902, https://doi.org/10.1002/jemt.20359.

[39] G.-G. Wang, M. Lu, Y.-Q. Dong, X.-J. Zhao, Self-adaptive extreme learning machine, Neural Comput. Appl. 27 (2) (2016) 291–303, https://doi.org/10.1007/s00521-015-1874-3.

[40] J.-H. Yi, J. Wang, G.-G. Wang, Improved probabilistic neural networks with self-adaptive strategies for transformer fault diagnosis problem, Adv. Mech. Eng. 8 (1) (2016), 1687814015624832, https://doi.org/10.1177/1687814015624832.

[41] S. Kosov, K. Shirahama, C. Li, M. Grzegorzek, Environmental microorganism classification using conditional random fields and deep convolutional neural networks, Pattern Recogn. 77 (2018) 248–261, https://doi.org/10.1016/j.patcog.2017.12.021.

[42] C. Li, K. Shirahama, M. Grzegorzek, Application of content-based image analysis to environmental microorganism classification, Biocybern. Biomed. Eng. 35 (1) (2015) 10–21, https://doi.org/10.1016/j.bbe.2014.07.003.

[43] J. Zhang, C. Li, S. Kosov, M. Grzegorzek, K. Shirahama, T. Jiang, C. Sun, Z. Li, H. Li, Lcu-net: a novel low-cost u-net for environmental microorganism image segmentation, Pattern Recogn. 115 (2021), https://doi.org/10.1016/j.patcog.2021.107885, 107885.

[44] F. Kulwa, C. Li, X. Zhao, B. Cai, N. Xu, S. Qi, S. Chen, Y. Teng, A state-of-the-art survey for microorganism image segmentation methods and future potential, IEEE Access 7 (2019) 100243–100269, https://doi.org/10.1109/ACCESS.2019.2930111.

[45] C. Sun, C. Li, J. Zhang, M.M. Rahaman, S. Ai, H. Chen, F. Kulwa, Y. Li, X. Li, T. Jiang, Gastric histopathology image segmentation using a hierarchical conditional random field, Biocybern. Biomed. Eng. 40 (4) (2020) 1535–1555, https://doi.org/10.1016/j.bbe.2020.09.008.

[46] Z. Cui, F. Xue, X. Cai, Y. Cao, G.-g. Wang, J. Chen, Detection of malicious code variants based on deep learning, IEEE Trans. Ind. Inf. 14 (7) (2018) 3187–3196, https://doi.org/10.1109/TII.2018.2822680.
[47] M. Shen, C. Li, W. Huang, P. Szyszka, K. Shirahama, M. Grzegorzek, D. Merhof, O. Deussen, Interactive tracking of insect posture, Pattern Recogn. 48 (11) (2015) 3560–3571, https://doi.org/10.1016/j.patcog.2015.05.011.
[48] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR), 2017, pp. 936–944, https://doi.org/10.1109/CVPR.2017.106.
[49] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904, https://doi.org/10.1109/TPAMI.2015.2389824. –1916.
[50] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR), 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.
[51] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, https://doi.org/10.48550/arXiv.1409.1556, 2015.
[52] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR), 2016, pp. 2818–2826, https://doi.org/10.1109/CVPR.2016.308.
[53] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, D. Ren, Distance-iou loss: faster and better learning for bounding box regression, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 12993–13000, https://doi.org/10.1609/aaai.v34i07.6999.
[54] A. Chen, C. Li, S. Zou, M.M. Rahaman, Y. Yao, H. Chen, H. Yang, P. Zhao, W. Hu, W. Liu, G. Marcin, Svia dataset: a new dataset of microscopic videos and images for computer-aided sperm analysis, Biocybern. Biomed. Eng. 42 (1) (2022) 204–214, https://doi.org/10.1016/j.bbe.2021.12.010.
[55] Y. Hu, J. Lu, Y. Shao, Y. Huang, N. Lü, Comparison of the semen analysis results obtained from two branded computer-aided sperm analysis systems, Andrologia 45 (5) (2013) 315–318, https://doi.org/10.1111/and.12010.
[56] I. Loshchilov, F. Hutter, Sgdr: stochastic gradient descent with warm restarts, arXiv preprint arXiv:1608.03983, https://doi.org/10.48550/arXiv.1608.03983, 2017.
[57] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, M. Pietikäinen, Deep learning for generic object detection: a survey, Int. J. Comput. Vis. 128 (2) (2020) 261–318, https://doi.org/10.1007/s11263-019-01247-4.
[58] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149, https://doi.org/10.1109/TPAMI.2016.2577031.
[59] M. O'connell, N. Mcclure, S. Lewis, The effects of cryopreservation on sperm morphology, motility and mitochondrial function, Hum. Reprod. 17 (3) (2002) 704–709, https://doi.org/10.1093/humrep/17.3.704.
[60] G.-G. Wang, S. Deb, Z. Cui, Monarch butterfly optimization, Neural Comput. Appl. 31 (7) (2019) 1995–2014, https://doi.org/10.1007/s00521-015-1923-y.
[61] G.-G. Wang, S. Deb, L.D.S. Coelho, Earthworm optimisation algorithm: a bio-inspired metaheuristic algorithm for global optimisation problems, Int. J. Bio-Inspired Comput. 12 (1) (2018) 1–22, https://doi.org/10.1504/IJBIC.2018.093328.
[62] G.-G. Wang, S. Deb, L.d.S. Coelho, Elephant herding optimization, in: 2015 3rd International Symposium on Computational and Business Intelligence, ISCBI), 2015, pp. 1–5, https://doi.org/10.1109/ISCBI.2015.8.
[63] J. Li, H. Lei, A.H. Alavi, G.-G. Wang, Elephant herding optimization: variants, hybrids, and applications, Mathematics 8 (9) (2020) 1415, https://doi.org/10.3390/math8091415.
[64] G.-G. Wang, Moth search algorithm: a bio-inspired metaheuristic algorithm for global optimization problems, Memetic Comput. 10 (2) (2018) 151–164, https://doi.org/10.1007/s12293-016-0212-3.
[65] S. Li, H. Chen, M. Wang, A.A. Heidari, S. Mirjalili, Slime mould algorithm: a new method for stochastic optimization, Future Generat. Comput. Syst. 111 (2020) 300–323, https://doi.org/10.1016/j.future.2020.03.055.
[66] Y. Yang, H. Chen, A.A. Heidari, A.H. Gandomi, Hunger games search: visions, conception, implementation, deep analysis, perspectives, and towards performance shifts, Expert Syst. Appl. 177 (2021) 114864, https://doi.org/10.1016/j.eswa.2021.114864.
[67] W. Li, G.-G. Wang, A.H. Gandomi, A survey of learning-based intelligent optimization algorithms, Arch. Comput. Methods Eng. 28 (5) (2021) 3781–3799, https://doi.org/10.1007/s11831-021-09562-1.
[68] J. Tu, H. Chen, M. Wang, A.H. Gandomi, The colony predation algorithm, JBE 18 (3) (2021) 674–710, https://doi.org/10.1007/s42235-021-0050-y.
[69] M. Li, G.-G. Wang, A review of green shop scheduling problem, Inf. Sci. 589 (2022) 478–496, https://doi.org/10.1016/j.ins.2021.12.122.
[70] A.A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, H. Chen, Harris hawks optimization: algorithm and applications, Future Generat. Comput. Syst. 97 (2019) 849–872, https://doi.org/10.1016/j.future.2019.02.028.