2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-7358-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISBI53787.2023.10230550

TOD-NET: TRANSFORMER-BASED NEURAL NETWORK FOR TINY OBJECT DETECTION IN SPERM MICROSCOPIC VIDEOS

J. Zhang[1,8], S. Zou[1], C. Li[1,], Y. Yao[2], M. Rahaman[3], W. Qian[4], H. Sun[5], M. Grzegorzek[6] and G. Wang[7,*]*

[1] College of Medicine and Biological Information Engineering, Northeastern University, China
[2] Department of Electrical and Computer Engineering, Stevens Institute of Technology, US
[3] School of Computer Science and Engineering, University of New South Wales, Australia
[4] The University of Texas at El Paso, US
[5] Shengjing Hospital, China Medical University, China
[6] Institute of Medical Informatics, University of Luebeck, Germany
[7] Department of Biomedical Engineering, Rensselaer Polytechnic Institute, US
[8] College of Electronic Science, National University of Defense Technology, China
[*] Corresponding author: C. Li, lichen@bmie.neu.edu.cn; G. Wang, wangg6@rpi.edu

ABSTRACT

The total number of families who have lost their only child in China is about 1 million, and the death toll in this category is about 76,000 yearly. Therefore, people desperately need the help of in vitro fertilization (IVF) technology, and the selection of excellent sperms is the key application of IVF technology. However, there exists some difficulties to detect tiny objects such as sperms in microscopic videos, especially in large-scale high-throughput experiments. One of the primary reasons is, sperms in microscopic videos are tiny, fuzzy, and with quite random characteristics and dynamics, which are difficult to detect using the current image analysis methods. Here, an advanced transformer-based neural network is proposed for tiny object detection (TOD-Net), and the model is evaluated on a unique high-quality labeled big dataset of sperm microscopic videos (consisting of >151,000 annotated objects). The results show that TOD-Net outperforms the state-of-the-art methods in the sperm detection task (83.61% AP_{50}), works in real-time (35.7 frames per second), and is in an excellent agreement with that reported by medical expert.

Index Terms— Image analysis, deep learning, tiny object detection, microscopic videos, sperm analysis

1. INTRODUCTION

Sperms are necessary for human and mammal reproductive processes. According to federal data, the total number of families who have lost their only child in China is about 1 million, and the death toll in this category is about 76,000 yearly [1]. Therefore, they desperately need help from IVF technology to select excellent sperm. High-quality sperms are also crucial in animal husbandry [2, 3, 4]. Hence, the detection and analysis of sperms play an essential role in human and animal breeding, especially in vitro fertilization (IVF) [5]. As the first module of the current Computer Aided Semen Analysis (CASA) system, sperm detection impacts the accuracy and reliability of the final analysis results. Currently, most sperm detection techniques [6, 7] are based on traditional image processing techniques such as thresholding, edge detection, and contouring. However, with these techniques, the detection results require manual intervention and often fail to distinguish healthy sperm and similar impurities. The typical difficulties in sperm detection are mainly due to small or tiny sizes of sperms, their diverse and fuzzy morphologic features, low contrast, the similarity between sperms and impurities, dense populations, which are shown in Fig. 1.

In recent several years, increasingly more object detection models are constantly being proposed in the artificial intelligence (AI) framework, such as Region-based CNN (RCNN) [8], You Only Look Once (YOLO) [9], Single Shot Multibox Detector (SSD) [10], and others [11]. The performance of Convolutional Neural Networks (CNNs) and Transformers has now surpassed the complex classic image processing algorithms in the task of object detection [12], which motivates us to develop deep learning methods for real-time sperm detection in microscopic videos. However, the accuracy of sperm object detection is still much lower than daily object detection under normal conditions [13]. A primary reason for the performance gap mentioned above is the lack of a publicly available annotated extensive dataset [14, 15]. Through five years of hard work, a unique dataset called MIaMIA-SpermVideo was developed, consisting of 111 long videos with more than 151,000 objects for sperm detection, tracking, and classification [16]. When annotating the MIaMIA-SpermVideo dataset, the sperms and impurities are meticulously marked so that TOD-Net can lever-

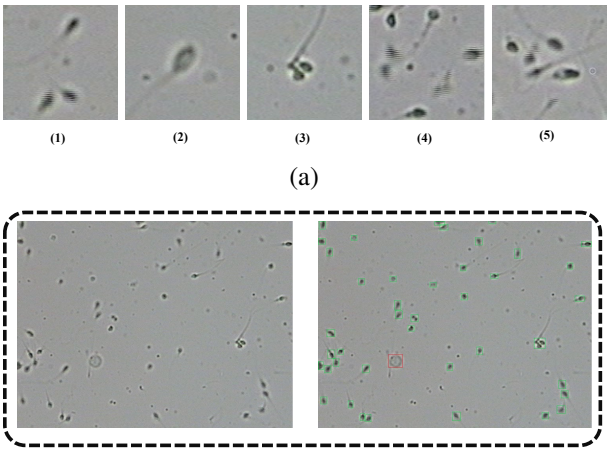

Fig. 1. Challenging cases for detection of sperms and examples of the datasets. In (a): (1) Normal sperms, (2-5) abnormal sperms and impurities; In (b): The sperm microscopic video (left) and the annotation for object detection (right). The green and red boxes represent the detected sperms and impurities, respectively.

age this knowledge to distinguish sperms from impurities. The details on the MIaMIA-SpermVideo dataset, TOD-Net codes, and system demo videos are all available at https://github.com/Demozsj/Detection-Sperm. We welcome researchers to share more data with us for further evaluation and enhancement of TOD-Net.

2. TOD-NET BASED SPERM DETECTION METHOD

The TOD-Net model, developed based on the YOLO model and transformers, can be regarded as a regression solver for fast and precise detection of small/tiny objects. The overall network architecture of TOD-Net is shown in Fig. 2(a), and the implementation details are presented as follows.

In Fig. 2(d), the coordinates of the predicted box are computed as the network output based on priori boxes, where t_x, t_y, t_w and t_h are predicted by TOD-Net, P_0 and P_1 represent the probability of being a sperm or an impurity in the bounding box respectively, C is the confidence regarding whether there is an object in the bounding box.

The object detection in video streams is one of the essential image processing tasks. FIrstly, the sperm microscopic videos are split into continuous frames. However, there exits some blurred frames due to the movement of the lens when capturing the sperm microscopic videos. After that, the blurred and normal frames can be distinguished based on the grayscale histogram analysis, and the blurred frames can be deleted based on the discrimination of the Otsu method. Furthermore, the proposed TOD-Net is an anchor-based object

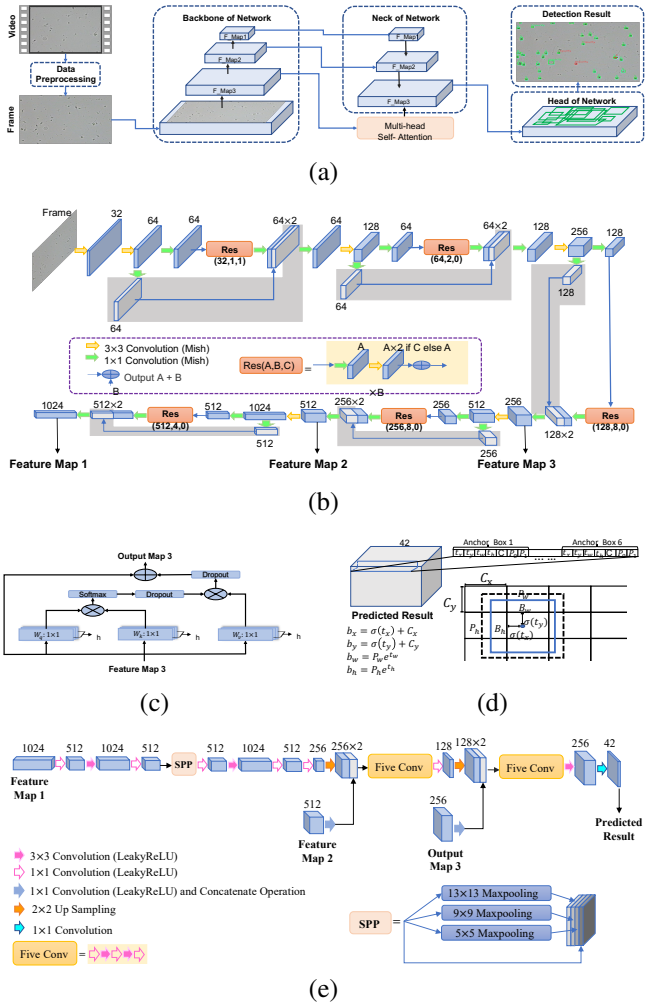

Fig. 2. Architecture of the proposed TOD-Net model. (a) The architecture of TOD-Net; (b) TOD-Net backbone; (c) Multi-head Self-Attention module; (d) TOD-Net head; (e) TOD-Net neck.

detection model, hence, the k-means algorithm is applied for frames clustering into a specific number of categories, which is set as six.

A straightforward backbone structure with cross-layer concatenate operations is shown in Fig. 2(b). The proposed approach focuses on detecting tiny objects with their accurate localization at once. First, inspired by ResNet, the Shortcut Connection is applied with a cross-layer addition operation in TOD-Net (as shown in Res (A, B, C) in Fig. 2(b)). Then, a cross-layer concatenation operation based on the straightforward backbone structureis is introduced to facilitate the information transfer (indicated by the grey arrows in Fig. 2(b)).

The proposed multi-head self-attention (MHSA) mechanism is shown in Fig. 2(c). Since the part of feature information of tiny objects (sperm) is lost during the the down-

sampling operation in the convolution layers, which may frustrate the detection results of tiny object. Therefore, a multi-head self-attention mechanism that excels in various image analysis tasks is introduced. The multi-head self-attention mechanism captures essential information in the feature map from multiple angles in parallel and builds the global and local connections to highlight small objects of interest in the feature map.

The model neck is applied to integrate features extracted from the backbone for object detection. The architecture of TOD-Net neck is shown in Fig. 2(f). In fact, there exists several abnormal sperms (very big) and other impurities (such as bacteria, protein lumps, and bubbles) in semen. These sperms and impurities are significantly different from normal sperms. Therefore, the spatial pyramid pooling [17] is performed to synergize multi-scale information for better analysis.

The objects with tiny size are easily lost while perfroming down-sampling operation. Hence, a feature fusion method is applied in the TOD-Net neck to solve the problem of object lost, where the shallow and deep feature maps are combined via up-sampling. In the head, 6 bounding boxes are predicted per 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 computed. For each cell, the offset from the upper left corner of the image is denoted as (C_x, C_y), and the width and height of the corresponding a priori box are denoted as P_w and P_h. The center coordinates (b_x and b_y), width (b_w) and height (b_h) of the predicted box are then calculated, as shown in Fig. 2(d). Multi-label classification is performed to predict the categories associated with the bounding boxes. Furthermore, the dense prediction is applied by the TOD-Net head, and the maximum value suppression [18] is enforced to remove bounding boxes with high overlap in the network output.

3. EXPERIMENT

3.1. Data Processing

All the sperm microscopic videos were recorded by a WLJY-9000 Computer-aided Sperm Analysis System [18] in a reproductive health hospital (Jinghua Hospital, Shenyang, China). More than 151,000 objects (sperms and impurities) were annotated, as illustrated in Fig. 1(b). Fourteen reproductive doctors and biomedical scientists helped to annotate the objects by using the Labeling software [19]. The well-curated dataset contains more than 125,000 objects from 101 long videos within bounding boxes with categorical information for tiny object detection.

3.2. Model Training and Evaluation

The task of object detection consists of positioning and classification steps. The complete intersection over union [20] (CIoU) function (location loss) and the binary cross-entropy function (confidence and classification loss) are applied as the loss function of the network, which are optimized using the Adam optimizer. Besides, the cosine annealing scheduler [21] was applied to adaptively adjust the learning rate. To quantitatively compare various object detection methods, popular metrics were applied here, including Recall (Rec), Precision (Pre), F1 Score (F1), and Average Precision (AP). Finally, the performance metrics were calculated (AP, F1, Pre, and Rec) using the codes from https://github. com/Cartucho/mAP [22].

Besides, a more suitable evaluation index was applied without affecting sperm positioning. This indicator considers a positive sample when the detected object meets the following two conditions at the same time: first, the detected object category is correct; and second, IoU of the detection box and the ground truth box exceeds B1 or the IoU of the detection box, the ground truth box exceeds B2, and the distance between the center points of the two boxes does not exceed R pixels.

4. RESULTS AND ANALYSIS

TOD-Net was systematically compared with several classical object detection models using the same datasets. As shown in Fig. 3(a), by comparing with the best results using various existing models, including YOLO-v3, YOLO-v4, SSD, RetinaNet, and Faster R-CNN, the AP is higher by 32.61, 40.68, 18.61, 68.56 and 47.85 perceptual points, respectively. The F1 score is higher by 28.42, 33.46, 20.39, 72.34 and 20.07 perceptual points, respectively. The Rec is higher by 30.79, 37.71, 19.47, 70.40 and 38.91 perceptual points, respectively. Though the Pre of TOD-Net is lower by 8 perceptual points than the best performing model (RetinaNet), but the Recall is higher by 72.34 perceptual points than RetinaNet, showing that the number of detected objects obtained by TOD-Net exceeds RetinaNet. Overall, TOD-Net outperforms the existing models in sperm detection. The detection results of these models are visualized in Fig. 3(c).

In Fig. 3(b), the APS, API and mAP represents AP of sperm, AP of impurity and mean AP, respectively. In Fig. 3(c), the blue boxes represent the ground-truth, the green boxes correspond to the correctly detected objects, and the red boxes indicate the incorrectly detected objects. The values under the images represent the number of true positive, false positive, false negative, and the total number of annotated objects, respectively.

Fig. 3(c) shows that the correct detection rate of TOD-Net is only slightly lower than that of Faster RCNN in "sperm-lack" scenes (oligospermia), but Pre is much higher than that of Faster RCNN [23]. In "sperm-normal" scenes (healthy), the correct detection rate of TOD-Net is optimal, and the Pre and Rec are higher than the other methods. Fig. 3(c) further explains why TOD-Net has slightly lower Pre than SSD and RetinaNet.

Moreover, 4,479 impurities were added into the dataset to test the robustness of TOD-Net. The results are shown in

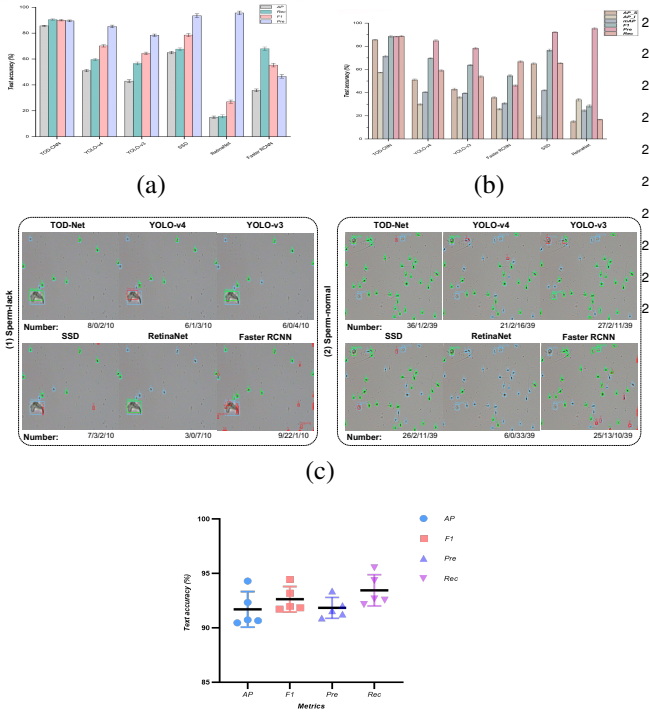

Fig. 3. Experimental results. (a) The comparison of different models; (b) the comparison of different models in the scene with impurity; (c) comparison of visualization result; (d) the detection results in the five-fold cross-validation process;

Fig. 3(b), where TOD-Net showed the best robustness against the impurities compared to the other models.

Furthermore, a five-fold cross-validation was applied to assess the repeatability of TOD-Net. The experimental results are shown in Fig. 3(d), where the mean values (μ) of the four evaluation metrics are higher than 91%. The standard deviation (STDEV) of Pre is 0.86%, the STDEV of AP is 1.47%, and the STDEV values of the rest two evaluation metrics are below 1.30%, which demonstrate that TOD-Net is very stable.

5. DISCUSSION

As shown in Fig. 4(a), the detection boxes often cover 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 the detection and ground truth boxes to be lower than 0.5. In Fig. 4(b), due to the movement of the sperms, the thickness of the semen wet film, and noticeable interference fringes in sperm videos, the important morphological information may be lost, leading to detection errors. Also, because impurities are visually similar to sperms, TOD-Net may incorrectly report impurities as sperms. In Fig. 4(c), the sperms on figure edges are difficult to process, and sometimes are missed in detection. To overcome this problem, only the sperms without controversy are labelled. In Fig. 4(d), the sperms may be deeply in the semen wet film, and it could be difficult to distinguish whether it is a sperm or an impurity. In that case, it was not annotated in our dataset. Besides, though the detection precision of sperms exceeded all existing networks, the precision of impurities is relatively unsatisfactory. Two perhaps reasons are those impurities vary in shape, and the quantity of the labeled impurities is limited.

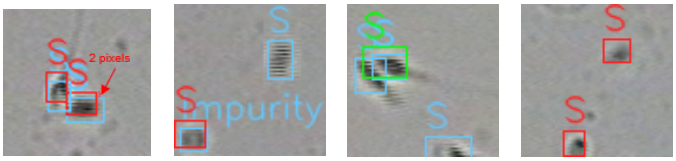

Fig. 4. Visualization of typical detection failures using TOD-Net. The red and green boxes represent the detection results, the blue boxes represent the ground truth, where S denotes sperms.

In conclusion, a high-quality massive dataset is developed for sperm detection, tracking, and classification. Also, an end-to-end CNN model (TOD-Net) for tiny object detection in real-time was developed and validated, which can accurately detect sperms in videos and images. Overall, the proposed TOD-Net has improved the performance of sperm image analysis, outperformed the existing methods. In the future, the relationship between the front and back frames in sperm microscopic video can be extracted and analysed for better detection results. Besides, the proposed tiny object detection network can also be applied to detect cells in flow cytometry analysis in the future [24, 25].

6. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by (Source information). Ethical approval was not required as confirmed by the license attached with the open access data.

7. ACKNOWLEDGMENTS

Jiawei Zhang and Shuojia Zou contribute equally to this paper and should be considered as co-first author. This work was supported by the "National Natural Science Foundation of China" (No. 82220108007).

8. REFERENCES

[1] Y. Song, "Losing an Only Child: the One-child Policy and Elderly Care in China," *Reproductive Health Matters*, vol. 22, no. 43, pp. 113–124, 2014.

[2] M. Manafi, *Artificial Insemination in Farm Animals*, BoD–Books on Demand, 2011.

[3] C. Keefer, "Artificial Cloning of Domestic Animals," *Proceedings of the National Academy of Sciences*, vol. 112, no. 29, pp. 8874–8878, 2015.

[4] M. You, "Changes of China's Regulatory Regime on Commercial Artificial Breeding of Terrestrial Wildlife in Time of COVID-19 Outbreak and Impacts on the Future," *Biological Conservation*, vol. 250, pp. 108756, 2020.

[5] S. Gadadhar, G. Viar, J. Hansen, and et al., "Tubulin Glycylation Controls Axonemal Dynein Activity, Flagellar Beat, and Male Fertility," *Science*, vol. 371, no. 6525, pp. eabd4914, 2021.

[6] V. Abbiramy, V. Shanthi, and C. Allidurai, "Spermatozoa Detection, Counting and Tracking in Video Streams to Detect Asthenozoospermia," in *Proc. of ICSIP 2010*, 2010, pp. 265–270.

[7] Z. Lu, X. Zhang, C. Leung, and et al., "Robotic ICSI (Intracytoplasmic Sperm Injection)," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 7, pp. 2102–2108, 2011.

[8] R. Girshick, J. Donahue, T. Darrell, and et al., "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *Proc. of CVPR 2014*, 2014, pp. 580–587.

[9] J. Redmon, S. Divvala, R. Girshick, and et al., "You Only Look Once: Unified, Real-time Object Detection," in *Proc. of CVPR 2016*, 2016, pp. 779–788.

[10] W. Liu, D. Anguelov, D. Erhan, and et al., "Ssd: Single Shot Multibox Detector," in *Proc. of ECCV 2016*, 2016, pp. 21–37.

[11] J. Zhang, P. Ma, T. Jiang, and et al., "SEM-RCNN: A Squeeze-and-excitation-based Mask Region Convolutional Neural Network for Multi-class Environmental Microorganism Detection," *Applied Sciences*, vol. 12, no. 19, pp. 9902, 2022.

[12] Z. Zou, Z. Shi, Y. Guo, and et al., "Object Detection in 20 Years: A Survey," *arXiv: 1905.05055*, 2019.

[13] J. Li, X. Liang, Y. Wei, and et al., "Perceptual Generative Adversarial Networks for Small Object Detection," in *Proc. of CVPR 2017*, 2017, pp. 1222–1230.

[14] X. Li, C. Li, F. Kulwa, and et al., "Foldover Features for Dynamic Object Behaviour Description in Microscopic Videos," *IEEE Access*, vol. 8, pp. 114519–114540, 2020.

[15] S. Zou, C. Li, H. Sun, and et al., "TOD-CNN: An Effective Convolutional Neural Network for Tiny Object Detection in Sperm Videos," *Computers in Biology and Medicine*, vol. 146, pp. 105543, 2022.

[16] A. Chen, C. Li, S. Zou, and et al., "SVIA Dataset: A New Dataset of Microscopic Videos and Images for Computer-aided Sperm Analysis," *Biocybernetics and Biomedical Engineering*, vol. 42, no. 1, pp. 204–214, 2022.

[17] K. He, X. Zhang, S. Ren, and et al., "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[18] Y. Hu, J. Lu, Y. Shao, and et al., "Comparison of the Semen Analysis Results Obtained from Two Branded Computer-aided Sperm Analysis Systems," *Andrologia*, vol. 45, no. 5, pp. 315–318, 2013.

[19] T. Lin, "LabelImg," *Online: https://github.com/tzutalin/labelImg*, 2015.

[20] Z. Zheng, P. Wang, W. Liu, and et al., "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression," in *Proc. of AAAI-CAI 2020*, 2020, pp. 12993–13000.

[21] I. Loshchilov and F. Hutter, "Sgdr: Stochastic Gradient Descent with Warm Restarts," *arXiv: 1608.03983*, 2016.

[22] J. Cartucho, R. Ventura, and M. Veloso, "Robust Object Recognition through Symbiotic Deep Learning in Mobile Robots," in *Proc. of IOOS 2018*, 2018, pp. 2336–2341.

[23] S. Ren, K. He, R. Girshick, and et al., "Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[24] J. Zhang, C. Li, M. Rahaman, and et al., "A Comprehensive Review of Image Analysis Methods for Microorganism Counting: from Classical Image Processing to Deep Learning Approaches," *Artificial Intelligence Review*, pp. 1–70, 2022.

[25] J. Zhang, C. Li, M. Rahaman, and et al., "A Comprehensive Survey with Quantitative Comparison of Image Analysis Methods for Microorganism Biovolume Measurements," *Archives of Computational Methods in Engineering*, pp. 1–35, 2022.