Automated Non-Invasive Analysis of Motile Sperms Using Sperm Feature-Correlated Network

Wei Dai, *Graduate Student Member, IEEE*, Zixuan Wu, Rui Liu, Tianyi Wu, Min Wang, *Graduate Student Member, IEEE*, Junxian Zhou, Zhuoran Zhang, *Member, IEEE*, and Jun Liu, *Senior Member, IEEE*

***Abstract*— An unbiased assessment of sperm morphology and motility is crucial for assessing fertility potential and guiding visual feedback for microrobotic manipulation. Automated analysis and selection of optimal sperm are essential for in vitro fertilization treatments, such as robotic intracytoplasmic sperm injection. However, conventional image processing methods face limitations in analyzing small sperm objects under microscopic imaging. While convolutional neural networks (CNNs) have brought promising advancements in microscopic image analysis, previous CNN methods have struggled to accurately differentiate tiny objects. These methods often require staining or fluorescence techniques to enhance visual contrast between sperm and culture medium, leading to clinical impracticality. To address these limitations, we introduce a novel sperm recognition network named the sperm feature-correlated network (SFCNet), for accurate and efficient segmentation and tracking of minute sperm objects. The SFCNet employs innovative modules, including collateral multi-scale convolution, cross-scale feature map guide, atrous spatial pyramid convolution with pooling, lateral attention, and multi-scale tracking proposal, to preserve essential sperm details despite their small size. Experimental results indicate that the SFCNet surpassed the state-of-the-art models designed for segmenting or tracking small objects, achieving up to a 28.39% higher Sørensen-Dice coefficient in segmentation and a 10.33% higher average precision in tracking. Additionally, the SFCNet excelled in sperm morphometric analysis, achieving errors below 15%. Moreover, the SFCNet also secured top-tier performance in sperm motility analysis, acquiring errors below 13% in seven sperm motility parameters.**

***Note to Practitioners*—This study is stimulated by the need to analyze the quality of motile sperms and select the optimal one for in vitro fertilization. Existing methods for detecting sperm fall short as they require a relatively high-magnification microscopic image or the usage of stain or fluorescence to increase sperm visualization, which limits the selection process or even makes the sperm clinically unavailable. To overcome these limitations, the present work proposes a new framework based on deep learning, which includes the design of extracting multi-scale sperm features. Experimental results suggest that the proposed method can perform better than existing methods in real-time analysis of multiple motile sperms' morphology and motility at 20× objective. In the future, there is a high potential for fertility specialists and healthcare workers to apply the presented framework in fertility treatment with higher accuracy and efficiency.**

***Index Terms*— Automation at micro/nano scale, sperm analysis, in vitro fertilization, deep learning, attention mechanisms.**

I. Introduction

INFERTILITY is a global health issue affecting millions of couples worldwide. Male factors alone account for 30% of infertility cases [1]. The morphology and motility of sperm are pivotal characteristics in determining its fertility potential and selecting healthy sperm for in vitro fertilization (IVF) in clinics.

Accurate measurement of sperm morphology and motility plays an essential role in evaluating sperm quality for addressing male infertility. The World Health Organization (WHO) has recommended key morphometric and motility parameters for assessing human sperm, encompassing head area, head length, head width, head ellipticity, head angle, tail length, VSL, VCL, VAP, ALH, MAD, LIN, WOB, and STR [2], as summarised in Fig. 1ab. Traditionally, subcellular analysis of sperm morphology and motility parameters has been conducted using high-magnification microscopy (100× objective) [3]. However, utilizing high magnification restricts the analysis to one sperm at a time due to the small field of view. To obtain an unbiased assessment of sperm morphology and motility across a semen sample and select the viable sperm from the population, it is necessary to evaluate multiple sperms under lower magnification microscopy (*e.g.*, 20× objective). However, a significant challenge arises as the area occupied by a single sperm is less than 1% of a petri dish under a 20× objective. Manual inspection and selection of small sperm cells are labor-intensive and necessitate extensive training for a physician to become proficient.

Recent advancements have focused on achieving precise localization of the sperm head center using the Kalman filter

Manuscript received 14 March 2024; accepted 19 May 2024. This article was recommended for publication by Associate Editor H. Lu and Editor L. Zhang upon evaluation of the reviewers' comments. This work was supported in part by the Research Grant Council (RGC) of Hong Kong under Grant 11217922, Grant 11212321, and Grant ECS-21212720; in part by Guangdong Province Basic and Applied Basic Research Fund Project under Grant 2019A1515110175; and in part by the Science and Technology Innovation Committee of Shenzhen under Grant Type-C SGDX20210823104001011. *(Corresponding author: Jun Liu.)*

Wei Dai, Zixuan Wu, Rui Liu, Tianyi Wu, Min Wang, Junxian Zhou, and Jun Liu are with the Centre for Robotics and Automation, City University of Hong Kong, Hong Kong, China (e-mail: jun.liu@cityu.edu.hk).

Zhuoran Zhang is with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen Campus, Shenzhen 518172, China.

Data is available on-line at: https://github.com/anthonyweidai/SFCNet.

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TASE.2024.3404488.

Digital Object Identifier 10.1109/TASE.2024.3404488

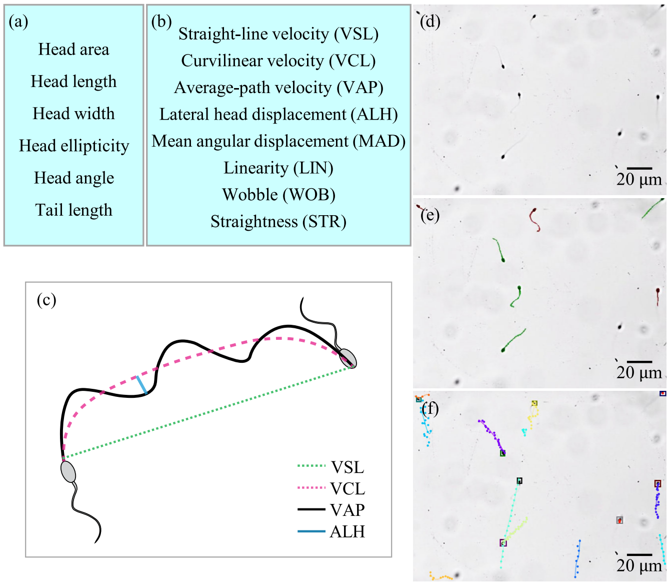

Fig. 1. Quantified parameters representing sperm (a) morphology and (b) motility. (c) Calculation of sperm motility parameters. (d) Exemplary semen image captured at 20× objective magnification. Automated (e) morphology and (f) motility analysis using computer vision algorithms.

for quantitative analysis of locomotive behaviour [4]. Additionally, Chang et al. [5] employed a k-means algorithm to detect color variations in culture dishes for identifying sperm heads. The watershed algorithm has also proven valuable in segmenting sperm from the surrounding medium from microscopic images [6]. Besides, Zhang et al. [7] applied a segmentation algorithm based on pixel intensity using a differential interference contrast (DIC) imaging mode to identify the sperm head, mid-piece, and tail. Despite their utility, these methods are ineffective in measuring a full spectrum of parameters for the assessment of sperm morphological structures and motility physics.

In addition to pixel-based image processing, deep learning algorithms were developed to recognize medical objects, including bacterial cells [8], blood cells [9], blastocysts [10], [11], polyps [12], [13], skin lesions [14], and so forth. Various deep learning methodologies, such as MobileNet [15] and UNet [16], have been utilized to analyze sperm characteristics. However, a common limitation of these methodologies lies in the necessity for fluorescent tags or staining dyes to enhance sperm visualization. Unfortunately, using foreign fluorochromes or dyes inevitably damages the cell health, making the sperm clinically impractical for IVF treatment. While Dai et al. successfully employed the UNet algorithm to accurately track individual sperm tails for robotic immobilization [17], and Liu et al. used the UNet-tiny model for non-invasive characterization of sperm head parameters [18], these approaches were limited to analyzing either individual sperm head [17] or tail [18] per instance. The non-invasive simultaneous measurement of both morphology and motility parameters for motile spermatozoa has remained largely unexplored.

In the field of computer vision, encoder-decoder architectures with a "U-shape" structure, such as UNet [19] and UNet++ [20]), have shown their potential in segmenting general objects. However, they often overlook the category of small objects. To tackle the challenge of reduced image resolution and information loss due to downsampling, feature correlation [21], [22], [23] and atrous convolution [24], [25] have been introduced. The spatial object contextual representation network (OCRNet) [21] leveraged the interaction among different-scale objects within an image. Additionally, deep labeling version three plus (DeepLabV3+) [24] and lite-reduced atrous spatial pyramid pooling (LRASPP) [25] expanded the receptive field by incorporating voids and captured broader context information through multi-scale context aggregation in atrous convolution, thereby eliminating the need for downsampling. The pyramid scene parsing network (PSPNet) [22] employed a multi-scale network to enhance the learning of global context representation, regardless of object sizes. Furthermore, the high-resolution network (HRNet) [23] kept high-resolution features in every layer of its architecture by using cross-resolution convolutions and information exchange.

For object tracking, recent advances have presented that multi-scale tracking yields promising results [26], [27], [28], [29], [30], particularly for small objects. The retina network (RetinaNet) [26] and fully-convolutional one-stage object detector (FOCS) [27] used pyramidal convolutions to produce the rich multi-scale features. Cascade region-based convolution neural network (Cascade RCNN) [28] applied a multi-stage strategy for reusing larger-shape tensors to maintain equivalence of detector quality and treat the same importance of objects regardless of their size. In addition, the trident network (TridentNet) [29] incorporated a parallel multi-branch system and sampled object instances based on their size. However, it remains unclear how effective these methods are in distinguishing spermatozoa.

This study proposes a novel deep learning architecture named the sperm feature-correlated network (SFCNet) to differentiate and characterize multiple sperms at a 20× objective magnification. SFCNet incorporates six core techniques: collateral multi-scale convolution, cross-scale feature map guide, atrous spatial pyramid convolution with pooling, lateral muti-scale attention, multi-scale region proposal, and joint probabilistic data association. Importantly, this methodology enabled morphology analysis (Fig. 1e) and motility analysis (Fig. 1f) without the need for fluorescence or dye staining to enhance sperm visibility. Experimental results demonstrate the superior performance of the SFCNet network, achieving a Dice score of 64.14% for segmenting sperm and errors of less than $<15\%$ across all measured morphology parameters. SFCNet also outperformed other tested methods with 92.10% AP_{50} and errors of less than $<13\%$ in measuring motility parameters on sperm tracking at a speed of 318 frames per second.

II. System Setup and Data Acquisition

This section presents the configuration of the microrobotic system in Sec. II-A and methods for processing and annotating the data in Sec. II-B.

A. System Setup

The system setup for the sperm analysis and manipulation was built on a standard inverted microscope

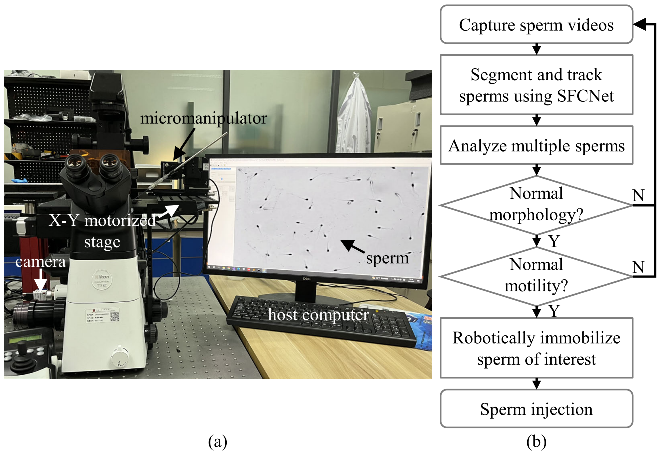

Fig. 2. (a) System setup for specimen collection and automated sperm analysis. (b) Flow diagram for analysis and selection of the sperm of interest.

(Nikon Eclipse Ti2), as depicted in Fig. 2a. A 20× objective lens (Nikon S Plan Fluor, NA: 0.45) was used for microscopic imaging. A CMOS camera (Basler A601f, with a dimension of 640×480) was used for capturing videos at 30 frames per second for analysis and visual feedback. A motorized 2-DOF translational stage (ProScan H117, Prior Scientific Inc.) was equipped to move the sperm on the X-Y plane. Advanced micromanipulation tasks such as sperm immobilization and injection were conducted with a 3-DOF micromanipulator (MP-285, Sutter Instrument Company) with a positioning resolution of $0.2\,\mu$m and a travel range of $25\,$mm for each axis. A host computer with an RTX3090 GPU paired with an Intel Xeon Platinum 8375C CPU was used to process the captured images and control the micromanipulation system.

B. Data Collection and Annotation

In this study, we collected semen samples from ten volunteers at the Prince of Wales Hospital in Hong Kong, preserving them using SpermCatch, a standard medium. All subjects provided consent forms in accordance with ethical protocols. The specimen images were extracted from the captured video clips at a sampling rate of one image for 15 frames. Subsequently, we constructed two distinct datasets, SpermSeg and SpermTrack, for the separate tasks of sperm segmentation and tracking. Following the WHO guidelines [2], experienced fertility doctors meticulously annotated the ground truth of the sperm entities in the images using the labelme [31]. To streamline the evaluation process, the datasets were divided into training and testing sets at a partition ratio of 4:1.

The SpermSeg dataset specifically comprises 148 images, with 118 used for training and 30 for testing. It includes two semantic classes: normal sperm (normal) and abnormal sperm (abnormal), annotated at the pixel level. The labeled dataset encompasses 618 instances of normal sperms, accounting for 42% of the total, and 852 instances of abnormal sperms, also representing 58% of the total. This amounts to a total of 1470 sperm instances. Given their small sizes, the sperm cells cover lower than 1% of the entire image area, approximately $0.042\% \sim 0.651\%$. Consequently, the non-sperm background occupies roughly 99% of the image.

Additionally, the SpermTrack dataset consists of 291 images, with 232 used for training and 59 for testing, and includes 3835 sperm objects. These objects are annotated at the box level. Given the difficulty in categorizing a sperm's motility characteristics within a single frame, the sperms in the SpermTrack dataset are not sorted. Instead, sperm motility was analyzed by post-processing the sequences of frames (refer to Sec. III-C3). The SpermTrack dataset includes more images and sperm instances than the SpermSeg dataset because annotating instances at the box level is less complex than at the pixel level.

III. Methodology

This section describes the key methodologies for automated sperm analysis with machine learning. The formulation and details of the sperm feature-correlated network (SFCNet) are explained in Sec. III-A.

A. Overall Deep Learning Framework

As depicted in Fig. 3, the SFCNet is composed of two primary elements: sperm segmentation (Sec. III-B) and sperm tracking (Sec. III-C). The segmentation component consists of four fundamental parts: collateral multi-scale convolution (Sec. III-B1), cross-scale feature correlation (Sec. III-B), atrous spatial pyramid convolution and pooling (Sec. III-B3), and sperm component measrement with muti-scale feature fusion (Sec. III-B4). Moreover, the tracking component of SFCNet includes three essential elements: Lateral attention with squeeze-excitation mechanism (Sec. III-C1), multi-scale tracking region proposal (Sec. III-C2), and joint probabilistic data association for sperm motility analysis (Sec. III-C3). Additionally, loss functions designed for sperm segmentation and tracking are discussed in Sec. III-B5 and Sec. III-C4, separately.

B. Cross-Scale Feature Guide for Segmentation

1) Collateral Multi-Scale Convolution: The network architecture comprises five horizontal stages. A bottleneck module (depicted by the pink arrow in Fig. 3) is applied to each stage. Each bottleneck unit consists of operations, including one 1×1 convolution followed by a 3×3 convolution and 1×1 convolution with a skip connection, functioning as a "bottleneck" in information theory. Assume that S_i represents the i^{th} stage, the dimension of the S_i is exactly $1/2^i$ of the dimension of the original input image. This progressive aggregation of features spans from lower to higher levels parallelly. Consequently, the feature maps encapsulate information from the proceeding stages to generate a comprehensive representation.

Between stages, a Conv3×3 is applied to downsize the feature maps and learn higher-level features (*e.g.*, shape, size, morphology, orientation, and motion behavior of sperms). All stages (*i.e.*, S_1, S_2, S_3, S_4, and S_5) function as the encoder component of the proposed segmentation model. Notably, various advanced backbones can be employed in the encoder part. In this study, the ResNet50 architecture [32] was chosen as the encoder.

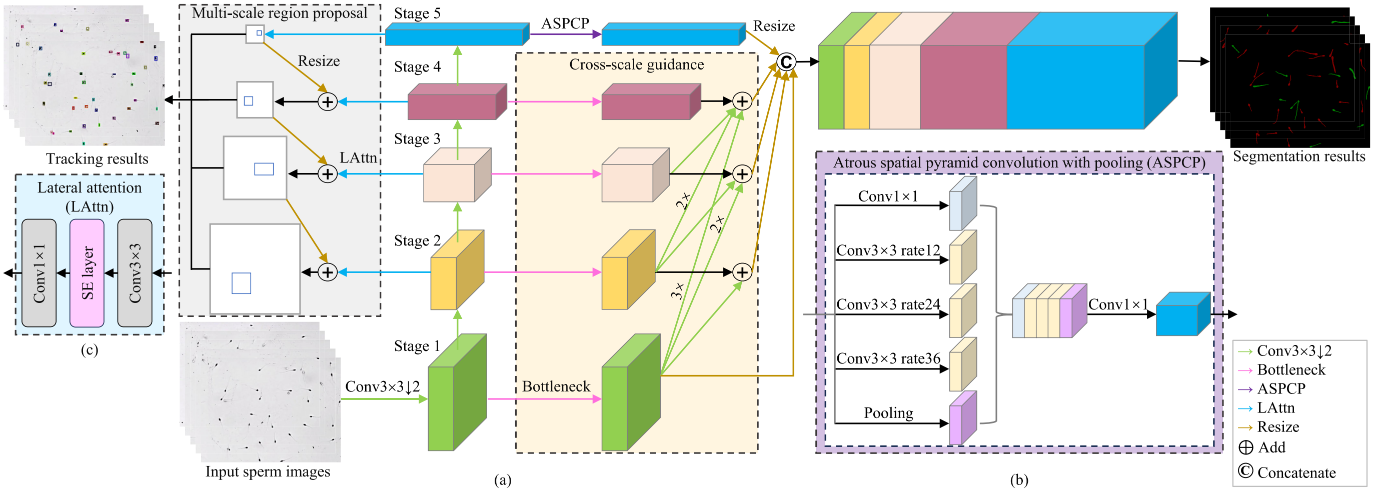

Fig. 3. (a) Architecture of sperm feature-correlated network. (b) Atrous spatial pyramid convolution with pooling in segmentation (refer to the right light-purple dashed box). (c) Lateral attention with the squeeze-excitation layer in tracking (refer to the left light-blue dashed box). The aggregation of low-level features with high-level features through strided convolution and lateral attention for segmenting and tracking sperms is illustrated in the middle light-yellow dashed box and the left light-grey dashed box, respectively.

2) Cross-Scale Feature Correlation: The intrinsic challenge for detecting small objects lies in preserving the feature of diminutive entities such as sperms during a sequence of convolutions using a stride of 2 (strided convolution). Hence, it is imperative to leverage the potential of features characterized by larger dimensions in the initial stages. This study applies the cross-scale feature maps to guide subsequent stages in learning the representation of small objects like sperms.

An illustration of the feature map guidance spanning four distinct scales (stages) is highlighted in the middle light-yellow dashed box of Fig. 3. Within this context, given three input tensors, $\{R_i, i \in \{1, 2, 3\}\}$, the output tensor, R_j, $j \in \{2, 3, 4\}$, is calculated through the following equation:

$$R_j = f_{1,j}(R_1) + f_{2,j}(R_2) + f_{3,j}(R_3) \tag{1}$$

where the transformation $f_{i,j}(R_i)$ performs $(j-i)$ 3×3 strided convolutions in the i^{th} input stage and j^{th} output stage.

Because the final stage, S_5, is connected with an additional segmentation module that computes feature maps differently from the remaining stages, it is essential to note that cross-scale guidance is absent in S_5 unless explicitly stated.

3) Atrous Spatial Pyramid Convolution With Pooling: The final stage, S_5, is responsible for extracting the highest-level visual features of sperm entities. To enhance this process, an integral segmentation module is appended at the end of S_5. Taking inspiration from DeepLabV3 [33], which utilizes spatial pyramid pooling to capture multi-scale information from objects, we apply atrous spatial pyramid convolution with pooling (ASPCP) as the segmentation head in this study (refer to Fig. 3b). To ensure that the convolution can extract features across regions of varying sizes, one 1×1 convolution and three 3×3 atrous convolutions (refer to Fig. 4) with atrous rates 12, 24, and 36 are adopted. It is worth noting that when the rate is 1, the atrous convolution reverts to standard convolution. Additionally, the global average pooling is applied to incorporate global information into the model. The final five output tensors are then concatenated into a single tensor and fed into another 1×1 convolution. Since the regions of sampling information vary for the five operations, this approach can be viewed as a pyramid feature extraction strategy.

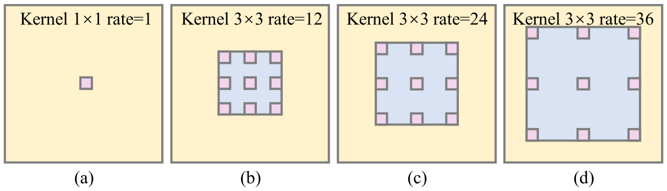

Fig. 4. Atrous spatial pyramid convolution of ASPCP. Atrous convolution with atrous rates (a) 1, (b) 12, (c) 24, and (d) 36 are applied to learn multi-scale features of sperms. Yellow, pink, and blue regions denote feature maps, convolution kernels, and convolution spaces, respectively.

4) Sperm Components Measurement: The outputs generated from five stages $(\text{S}_i, i = 1, 2, \ldots, 5)$ exhibit different feature scales. Therefore, a crucial step involves sampling these outputs to ensure uniform height and width dimensions. Since the output originating from S_1 has the dimension most similar to those of the original image, all sub-stream outputs are reshaped to align with the dimension of the S_1 output using the linear interpolation technique. The fusion of the multi-scale features combines all dimension levels of sperm characteristics to provide an accurate segmentation result. This process is visually depicted by the middle yellow arrows in Fig. 3.

Due to the relatively low resolution (96 DPI - dots per inch) of the image acquisition with the 20× objective and Basler A601f camera, the intricate morphology of the tiny sperm poses a challenge in accurate recognition. Considering these limitations, the analysis is focused on the following morphology parameters: head area, head length, head width, head ellipticity, head angle, and tail length. The automatic differentiation between the head and tail components of sperms is performed based on the distance between the component boundary and the skeleton of the sperm, as outlined in [34].

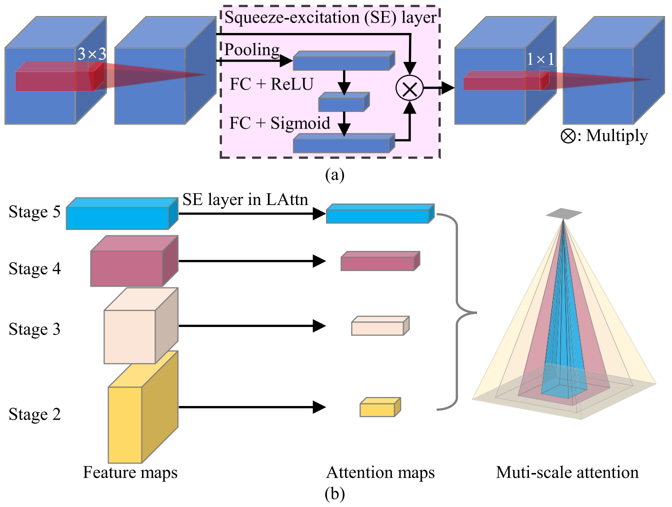

Fig. 5. (a) Lateral attention with squeeze-excitation layer. (b) Multi-scale attention through LAttn in different scales. FC is an acronym for a fully connected layer.

5) Loss Function: To quantify the discrepancy between the prediction mask and ground truth, we employed cross-entropy loss for the pixel-level classification in the segmentation task of SFCNet. The cross-entropy loss formula, $\mathcal{L}_{\mathrm{CE}}(p, c)$ is calculated as following:

$$\mathcal{L}_{\mathrm{CE}}(p, c) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c}\log_2(p_{n,c}) \tag{2}$$

where N represents the number of input images, C denotes the number of classes, $y_{n,c}$ is the binary indicator (0 for False, 1 for True), and $p_{n,c}$ is the predicted probability of the n^{th} image belonging to the c^{th} class.

C. Multi-Scale Attention for Tracking Proposal

1) Lateral Attention With SE Layer: To capture both global (*i.e.*, color distribution, object interaction, and texture of entire image) and local (*i.e.*, corner, edge, specific pattern, and color) representations of the sperm, a lateral attention (LAttn) mechanism with a squeeze-excitation (SE) layer is introduced. The LAttn mechanism consists of a 3×3 convolution, followed by an SE layer and 1 × 1 convolution. Two convolution operations are designed to learn the local features of sperms, while the SE layer is tasked with global features. The SE layer is adapted from [35]. The LAttn mechanism is presented in Fig. 3c and detailed in Fig. 5a.

As shown in Fig. 5b, the LAttn is utilized in four stages (*i.e.*, S_2, S_3, S_4, and S_5). The SE layers in LAttn can generate attention feature maps at multiple dimensions, functioning as multi-scale attention mechanisms. To enhance features with cross-scale information for small objects, the feature map from the lower stage is integrated with the upsampled feature map from the higher stages.

2) Muti-Scale Region Proposal: To generate bounding boxes for sperm objects, a considerable number of region proposals are introduced to encompass detected objects. However, as these proposals are generated randomly, this process can be both time- and resource-intensive. Inspired by Faster RCNN [36], we apply a small network with a 3×3 convolution to slide across the feature map. Furthermore, it is essential to incorporate larger-dimension features from lower network stages to bind an object as small as a spermatozoa effectively. Therefore, the merged feature maps from LAttn and upsampling at diverse stages are employed to produce proposals with hierarchical dimensions, as depicted in the light-grey dash box of Fig. 3. This multi-scale region proposal method produces bounding boxes of varying sizes, which can exceptionally localize the small sperm objects. This is particularly important as conventional region proposal networks may undervalue small object regions (*i.e.*, those occupying less than 1% of the area) [36], [37].

3) Sperm Motility Analysis: Given that the area occupied by a sperm head is significantly larger than that of its tail, the analysis of sperm motility is primarily conducted by tracking the head's movement. The center of the bounding box surrounding the sperm head is designated as the sperm's position. Sperm motility parameters are then calculated based on the sperm's trajectory, for example, VSL, which represents the velocity along the straight-line path. It is important to note that spermatozoa may overlap, leading to interpolated trajectories. To ensure accurate trajectory mapping of the target sperm, the joint probabilistic data association filter (JPDAF) [38] was utilized to associate trajectory points that belong to the same sperm.

Assume that the state of a sperm at frame t is represented as $[x_t, y_t, x_t', y_t']$. The Mahalanobis distance D_t, computed by JPDAF between the actual position $Q_t = [x_t, y_t]$ and predicted position $\tilde{Q}_t = [\tilde{x}_t, \tilde{y}_t]$, is given by:

$$D_t = \sqrt{(Q_t - \tilde{Q}_t)F^{-1}(Q_t - \tilde{Q}_t)} \tag{3}$$

where F represents the covariance matrix of the correct measurement.

Additionally, assume that the association probability between the predicted and actual positions is represented as $P_t(\alpha)$ for the α scenario. The total association probability, denoted as $\beta_{o,m}(t)$, is updated by traversing and summing up all scenarios for the o^{th} sperm and the m^{th} measurement. The total association probability is defined as follows:

$$\beta_{o,m}(t) = \sum_{\alpha} P_t(\alpha)\omega_{o,m}(\alpha, t) \tag{4}$$

where $P_t(\alpha)$ is inversely proportional to D_t, and $\omega_{o,m}(\alpha, t)$ equals 1 if sperm o is associated with measurement m in case α, and 0 otherwise.

4) Loss Function: The task of sperm tracking requires a regression loss for calculating bounding box coordinates and determining the category of the bounded object. Therefore, we employ smooth L_1 [39] and cross-entropy losses for sperm bounding box regression and classification, respectively. The smooth L_1 loss can be formulated as follows:

$$\mathcal{L}_{\mathrm{Reg}}(t^c, g) = \sum_{k\in\{\mathrm{x,y,w,h}\}} \mathcal{L}_1(t_k^c - g_k) \tag{5}$$

in which

$$\mathcal{L}_1(x) = \begin{cases} |x| - 0.5, & \text{if } |x| \geq 1 \\ 0.5\ x^2, & \text{otherwise} \end{cases} \tag{6}$$

where the ground truth coordinates are represented as $g_i = (g_{\mathrm{x}}, g_{\mathrm{y}}, g_{\mathrm{w}}, g_{\mathrm{h}})^i$ for class c, and the predicted values are represented as $t_k^c = (t_{\mathrm{x}}^c, t_{\mathrm{y}}^c, t_{\mathrm{w}}^c, t_{\mathrm{h}}^c)^k$.

The cross-entropy loss is the same as in Eq. (2), but the loss is calculated at the box-object level. The total tracking loss for SFCNet is then given by:

$$\mathcal{L}_{\mathrm{Det}} = [c \geq 1]\mathcal{L}_{\mathrm{Reg}}(t^c, g) + \mathcal{L}_{\mathrm{CE}}(p, c) \tag{7}$$

where $[c \geq 1]$ equals 1 when $c \geq 1$ (sperm), and 0 otherwise (background).

IV. Experimental Results

The proposed SFCNet architecture was evaluated and compared with the state-of-the-art (SOTA) machine learning algorithms in the experiments. The training and evaluating configurations are first introduced in Sec. IV-A. Subsequently, the visualization and analysis of the segmentation results are elucidated in Sec. IV-B, followed by an analysis of tracking results in Sec. IV-C. Finally, a comprehensive case study and ablation study are conducted and illustrated in Sec. IV-D and Sec. IV-E.

A. Implementation Details

1) Evaluation Metrics: For sperm segmentation, we employed five metrics for quantitative evaluation, including the Sørensen-Dice coefficient (Dice), mean intersection over union (mIoU), enhanced-alignment metric (E-measure) [40], sensitivity, and precision. Specifically, Dice, mIoU, E-measure, and sensitivity were applied to quantify the measurement, while precision was used to qualify the measurement.

For sperm tracking, we utilized three metrics for quantitative evaluation: average precision at 0.50:0.05:0.95 of mIoU (AP), average precision at 0.50 of mIoU (AP$_{50}$), and average precision at 0.75 of mIoU (AP$_{75}$). Since sperm tracking should be available in the actual application of microrobotic manipulation, frames per second (FPS), giga floating point operations per second (GFLOPS), and the number of parameters (# parameters) were measured to evaluate the time efficiency and computational cost.

The results of segmentation and tracking tasks were computed by averaging three separate training and testing cycles.

2) Other Configurations: For the sperm segmentation, the mini-batch size was set to 4. Pre-processing the input images involved random crop, resizing with a dimension of 512 × 512, Gaussian blur, distortion, and rotation. The optimization process employed the AdamW optimizer [41] and adopted cross-entropy loss and a 0.03 weight decay. The learning rate was adjusted using a cosine schedule [42], decreasing from 5×10^{-5} to 1×10^{-6}. Furthermore, the comprehensive training was carried out for 100 epochs.

For the sperm tracking task, the mini-batch size was set to 16, and the input images were resized to a dimension of 800×800. No additional data processing method was required.

TABLE I
Segmentation Results for Various Methods on the SpermSeg Dataset. Unit: %

Method	Dice ⇑	mIoU	E-measure	Sensitivity	Precision
OCRNet [21]	35.75	34.65	30.84	35.15	56.46
DeepLabV3+ [24]	47.76	41.49	71.79	45.06	57.37
CFANet [13]	54.22	45.83	92.79	53.36	59.08
UNet [19]	55.17	47.16	94.74	54.46	64.63
UNet++ [20]	55.43	47.34	93.74	54.71	65.21
PraNet [12]	56.26	47.26	91.63	53.48	61.41
HRNet [23]	59.95	50.39	97.13	58.49	63.43
SFCNet (ours)	**64.14**	**53.53**	**97.93**	**63.46**	**65.67**

The optimization process hinged on the SGD optimizer in conjunction with adopting smooth L1 and cross-entropy losses. The learning rate was adjusted using a multi-step schedule, decreasing from 0.02 to 2×10^{-5}. Additionally, comprehensive training was executed for 1000 iterations.

The codes were implemented by using the PyTorch [43] and Detectron2 [44] packages. The experimental computations were conducted on an RTX3090 GPU paired with an Intel Xeon Platinum 8375C CPU. All the aforementioned configurations of segmentation and tracking tasks are consistent across all tested methods. The encoder part of all tested networks was pretrained in the ImageNet-1K [45] dataset.

B. Segmentation Results and Analysis

1) Segmentation Evaluation: To evaluate the efficacy of the proposed method, eight SOTA tiny-object segmentation models (*i.e.*, OCRNet [21], DeepLabV3+ [24], CFANet [13], UNet [19], UNet++ [20], PraNet [12], and HRNet [23]) were included as reference points in the experiments.

As illustrated in Tab. I, the SFCNet yielded the highest Dice of 64.14%, mIoU of 53.53%, E-measure of 97.93%, sensitivity of 63.46%, and precision of 65.67%, outperforming the SOTA small object segmentation methods. For the coincidence degree between the detected region and ground truth, SFCNet delivered an improved performance of 4.19% ∼ 28.39% Dice and 3.14% ∼ 18.88% mIoU than SOTA methods, demonstrating that SFCNet is capable of locating the regions of sperm morphology effectively and provided a reliable visual signal for robotic cell surgery at micro-scale.

Furthermore, SFCNet also achieved up to 67.09% better E-measure among all tested models. Since E-measure simultaneously considers pixel-level (*i.e.*, region coincidence) and image-level (*i.e.*, noise and blur) errors, it can provide a comprehensive segmentation result on sperm recognition. The best E-measure attained by SFCNet has revealed superior robustness in recognizing sperm, regardless of image noise and blur.

Besides, SFCNet obtained 4.97% ∼ 28.31% better sensitivity and 2.24% ∼ 9.21% better precision compared to other tested models. Such results underscore that SFCNet can effectively perceive sperm by ensuring all sperms are likely to be detected, and the positively detected regions will likely cover sperms.

Additionally, the OCRNet obtained less than 40% Dice, E-measure, and sensitivity in the SpermSeg dataset, with

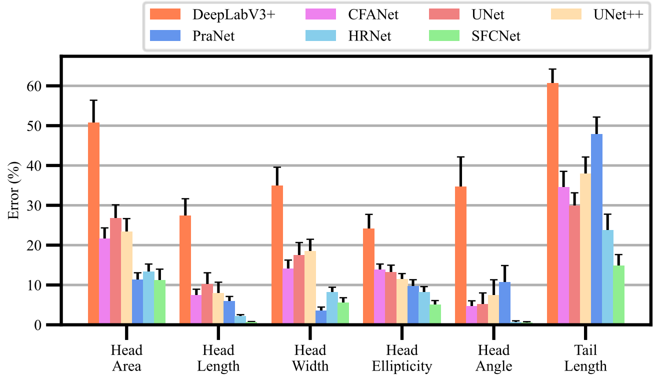

Fig. 6. Errors in automated morphometric analysis using top-7 deep learning methods compared to the manual benchmark.

TABLE II

TRACKING RESULTS FOR VARIOUS METHODS ON SPERMTRACK DATASET. FPS/GFLOPS USES THE ARBITRARY UNIT. "#" IS THE ABBREVIATION FOR "THE NUMBER OF"

Method	FPS/GFLOPS	# Parameters/M	AP_{50}/% ⇑	AP/%	AP_{75}/%
YOLOX [37]	335/185.30*	63.72*	82.87	29.81	11.74
FCOS [27]	386/185.02	32.24	83.20	31.53	12.37
RetinaNet [26]	339/215.53	37.92	84.81	33.05	13.16
Faster RCNN [36]	345/904.71	32.98	87.24	31.59	10.56
TridentNet [29]	144/594.90	32.98	90.88	37.50	17.88
Cascade RCNN [28]	145/211.01	69.10	91.53	38.82	20.57
SFCNet (ours)	318/249.30	49.56	**92.10**	**40.14**	**21.65**

*Reported by [37].

over a 25% gap of these metrics to SFCNet, suggesting that OCRNet hardly discriminated sperm from the background.

2) Sperm Morphology Analysis: To assess the performance of automated segmentation algorithms in the medical application of tested models, morphometric parameters were measured. The ground truth of 30 tested sperm images with 321 spermatozoa instances was measured using ImageJ by averaging annotation results from three independent expert technicians. The errors (± standard error) associated with automated quantification were assessed across various morphometric parameters using top-7 deep learning methods, DeepLabV3+, CFANet, UNet, UNet++, PraNet, HRNet, and SFCNet. These errors are summarised in Fig. 6. Since OCRNet struggled to differentiate sperms in the images and obtained more than 90% errors, OCRNet was not presented in Fig. 6.

The proposed SFCNet achieved the smallest errors in all sperm morphometric parameters, except for the head area, where it recorded the second-smallest errors. The error percentages ranged from 11.27±2.70% for head area, 0.65±0.17% for head length, 5.59±1.21% for head width, 5.14±0.94% for head ellipticity, 0.59±0.19% for head angle, and 14.89±2.74% for tail length. These results outperformed the second-best model, HRNet, by reducing averaged errors by up to -8.94%.

Notably, the errors in measuring sperm tails were the most pronounced among tested parameters. This phenomenon could be attributed to the relatively low DPI of 96 used in each image capturing using the 20× objective. Addressing this challenge could involve leveraging image enhancement techniques to boost the recognition of sperm tail features.

3) Visualization and Error Analysis: In addition to the quantitative evaluation, the segmentation results of the top-4 methods, UNet++, PraNet, HRNet, and SFCNet, were also compared with an in-depth sperm morphology analysis.

The prediction results of segmentation masks are exemplified in Fig. 7. It is clear from Fig. 7cd that UNet++ and PraNet struggled to recognize sperm locations, as evident from images without continuous masking. In other words, the detected sperm regions are disconnected or broken, as shown in the partially enlarged blue circles in Images No. 1 and 3 in Fig. 7cd. Conversely, although the second-best algorithm, HRNet, successfully recognized sperm continuous morphology in Fig. 7e, it fell short in precisely identifying sperm tails. Meanwhile, UNet++, PraNet, and HRNet ignored several sperm on the left-bottom part of Image 1. In contrast, the SFCNet exhibited the capability to accurately identify all sperm positions and effectively reconstruct the morphologies of sperms entities within the image (see Fig. 7f).

Furthermore, the sample sperm image has two classes, normal (negative) and abnormal (positive), represented by green and red regions in Fig. 7. Although UNet++, PraNet, and HRNet successfully detected the majority of sperm positions, they mistakenly categorized normal sperms as abnormal ones (false positive) or abnormal sperms as normal ones (false negative), as visually highlighted in Image No. 1 and 3 in Fig. 7c-e. However, the SFCNet proficiently differentiated normal and abnormal sperms, aligning closely with the ground truth, as evidenced by the green and red regions in Fig. 7bf.

In addition, several sperms may appear in the boundary of the receptive field. These sperms cannot be fully visualized, and only a partial head or tail can be observed. On the Image 2 purple circle region in Fig. 7, a sperm tail was annotated by doctors and another sperm tail was detected by PraNet and SFCNet, but it was not marked by the annotation. Moreover, it is difficult to examine a sperm, given partial morphology information. Because sperms are usually considered for actual medical applications, it is recommended to ignore those sperms whose bodies are not fully shown in the image.

C. Tracking Results and Analysis

1) Tracking Evaluation: To assess the effectiveness of the proposed method, six SOTA tiny-object tracking models, including YOLOX [37], FCOS [27], RetinaNet [26], Faster RCNN [36], TridentNet [29], and Cascade RCNN [28], were included as the control group in the experiments.

The tracking results on the SpermTrack dataset presented in Tab. II illustrate that SFCNet secured the first place in tracking sperms with 92.10% AP_{50}, 40.14% AP, and 21.65% AP_{75}. Besides, SFCNet performed at the speed of 318 FPS in tracking sperm, exceeding the requirements of real-time medical applications (*i.e.*, $>$ 30 FPS). Among all tested methods, SFCNet shows the competitive computational efficiency, 249.30 GLOPS, which is around 30% larger than the fastest method, FOCS, but achieved 8.90% better AP_{50}. Moreover, SFCNet has 49.56 M # parameters, smaller than the second-best tracking method, Cascade RCNN, with -19.44 M. These results highlight the superior performance of SFCNet in

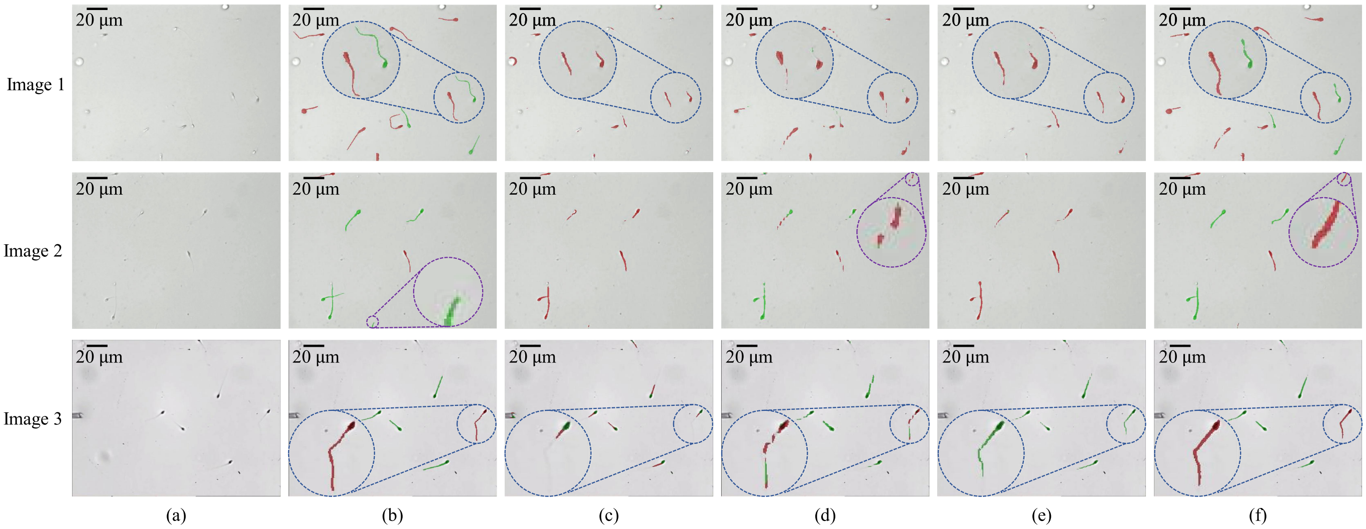

Fig. 7. Visualization of sperm segmentation results. (a) Original image and (b) ground truth. The others are detected results using top-4 methods: (c) UNet++, (d) PraNet, (e) HRNet, and (f) SFCNet (ours). Normal and abnormal sperms are covered by green and red colors, separately.

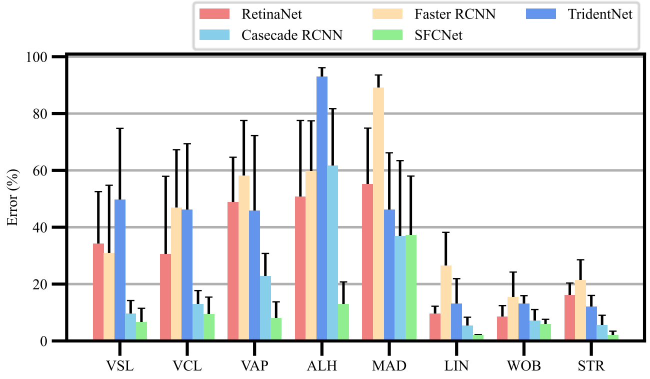

Fig. 8. Errors in automated motility analysis using the top-5 deep learning methods compared to the manual benchmark.

representation learning for sperm tracking and the computation efficiency of SFCNet in cell medical surgery and cell engineering.

2) Sperm Motility Analysis: To assess the performance of automated tracking algorithms in the medical application of tested models, the morphometric parameters were measured. Similar to Sec. IV-B2, we applied ImageJ to annotate the center points of sperm heads with 59 tested images by averaging annotation results from three independent expert technicians. The errors (± standard error) associated with automated quantification were assessed across various motility parameters. For more precise visualization, only top-5 deep learning methods, RetinaNet, Faster RCNN, TridentNet, Cascade RCNN, and SFCNet, are discussed, and their sperm tracking errors are summarised in Fig. 8.

The proposed SFCNet achieved the smallest averaged errors in all sperm motility parameters, except for MAD, where it recorded the second smallest averaged errors. The error percentages varied from 6.53±4.95% for VSL, 9.47±5.95% for VCL, 7.97±5.79% for VAP, 12.92±7.85% for ALH, 37.26±20.74% for MAD, 1.97±0.24% for LIN, 5.91±1.69% for WOB, and 2.03±1.41% for STR, outperforming the second-best model, Cascade RCNN, by an averaged reduction of −43.34%. Despite SFCNet recording the second smallest averaged errors in MAD, its highest MAD error was 5.42 % lower than that of the top-performing method, Cascade RCNN.

Remarkably, the errors in measuring MAD were the most pronounced among tested parameters. This phenomenon could be attributed to errors in the curve fitting for computing angular displacement. Addressing this challenge could involve using a camera to capture a higher image resolution.

3) Visualization and Error Analysis: In addition to the quantitative evaluation, the tracking results of the top-4 methods, Faster RCNN, TridentNet, Cascade RCNN, and SFCNet, were further compared for an in-depth sperm motility analysis.

The prediction results of tracking bounding boxes are revealed in Fig. 9. The results indicate that almost all sperm can be located by the top-4 methods, except TridentNet, which ignored one sperm in Image No. 1 (see first row in Fig. 9d). While Faster RCNN and Cascade RCNN can locate all sperm positions, they struggled to effectively identify the boundary of sperm head (see the first row in Fig. 9ce). Conversely, SFCNet successfully recognized sperm heads with the consistent bounding box as ground truth (see the first row in Fig. 9bf).

Unlike whole sperm segmentation in Sec. IV-B3, sperm tracking only focuses on the head of the sperm. Therefore, debris and dead sperm cells are more likely to be misclassified as sperm heads (false positive cases). For example, as presented in Images No. 2 and 3 in Fig. 9cd, Faster RCNN and TridentNet wrongly categorized debris and dead sperm cells as sperm heads. Meanwhile, Faster RCNN, TridentNet, and Cascade RCNN mistakenly estimated the micropipette as sperm, a false positive case that might cause medical accidents during surgery. In contrast, the SFCNet exhibited the capability to identify all sperm positions accurately and effectively without focusing on non-sperm objects (see Fig. 9f).

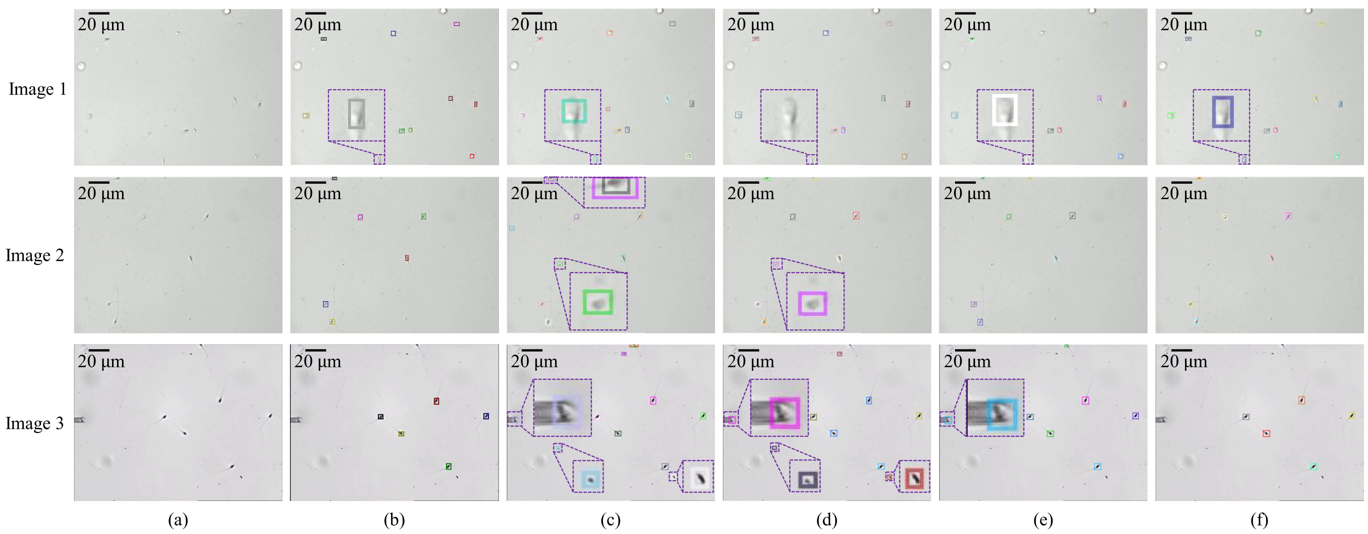

Fig. 9. Visualization of sperm tracking results. (a) Original image and (b) ground truth. The others are detected results using top-4 methods: (c) Faster RCNN, (d) TridentNet, (e) Cascade RCNN, and (f) SFCNet (ours). Each box proposal has a random and identical color.

TABLE III
AUTOMATED QUANTIFICATION OF SIX SPERM SAMPLES. AU: ARBITRARY UNIT. UNCOMMON DATA IS UNDERLINED

Sperm No.	Morphology							Motility									Healthy
	Head					Tail											
	area (μm^2)	length (μm)	width (μm)	ellipticity (AU)	angle (°)	length (μm)	Normal	VSL (μm/s)	VCL (μm/s)	VAP (μm/s)	ALH (μm/s)	MAD (°)	LIN (AU)	WOB (AU)	STR (AU)	Normal	
1	14.63	$\underline{8.49}$	2.92	$\underline{2.91}$	0.00	$\underline{26.35}$	✗	$\underline{3.95}$	$\underline{3.99}$	$\underline{4.02}$	0.27	2.74	0.99	1.01	0.98	✗	✗
2	12.50	5.50	3.00	1.83	0.00	44.24	✓	$\underline{6.81}$	$\underline{7.24}$	$\underline{6.67}$	1.16	2.60	0.94	0.92	1.02	✗	✗
3	13.63	4.95	3.54	1.40	0.00	38.60	✓	11.49	11.80	11.91	0.45	0.95	0.97	1.01	0.97	✓	✓
4	15.00	$\underline{8.01}$	2.50	$\underline{3.21}$	34.72	56.61	✗	14.71	16.93	15.67	0.48	0.07	0.87	0.93	0.94	✓	✗
5	14.25	$\underline{7.38}$	2.69	$\underline{2.74}$	2.76	36.18	✗	$\underline{3.74}$	$\underline{3.96}$	$\underline{3.77}$	0.48	9.47	0.95	0.95	0.99	✗	✗
6	18.00	$\underline{4.00}$	3.00	1.33	53.13	32.78	✗	$\underline{0.10}$	$\underline{0.46}$	$\underline{1.77}$	0.31	0.58	$\underline{0.22}$	$\underline{3.87}$	$\underline{0.06}$	✗	✗

Additionally, Faster RCNN provided multiple tracking box proposals for a sperm near the edge of Image No. 1 in Fig. 9c, giving the wrong signals that there were two sperms in the same place.

D. Case Study

To investigate the novel SFCNet method in the analysis of sperm morphological structures and motility, a case study was performed as a demonstration by randomly selecting six sperm samples. Subsequently, the predictive values of morphology and motility parameters computed by the SFCNet are presented in Tab. III. The calculated morphological parameter values for healthy sample sperms were 12.50∼13.63 μm^2 for head area, 4.95∼5.50 μm for head length, 3.00∼3.54 μm for head length, 1.40∼1.83 AU for head ellipticity, and 38.60∼44.24 μm for tail length. Moreover, Sperms No. 1, 4, and 5 exhibit more extensive head length than width, leading to abnormal head ellipticity ($>$2.7). Meanwhile, the tail length of the first sample fell below 30 μm, one of the characteristics of abnormal sperms. In addition, Sperm No. 6 has a comparatively small head length (4 μm) and ellipticity (1.33), which are considered abnormal sperm features.

In addition to morphology analysis, the motility measurement of sperm, as indicated in the right part of Tab. III, reveals that Sperms No. 1, 5, and 6 exhibited minimal movement, with VCL lower than 4.5 μm/s. Furthermore, Sperm No. 2 displayed 6.81 μm/s VSL, 7.24 μm/s VCL, and 6.67 μm/s VAP, indicating its weak motility characteristics. Consequently, Sperms No. 1, 2, and 6 are regarded as abnormal in terms of moving velocity. In contrast, Sperms No. 3 and 4 exhibited VCL and VAP exceeding 10 μm/s, and values of $\geq$0.87 for LIN, WOB, and STR, which fall within the normal range for sperm characteristics.

Furthermore, sperm No. 2 exhibited regular motility but had abnormal morphology. Thus, sperm No. 3 was identified by SFCNet as the only healthy sperm among the six samples.

E. Ablation Study

1) Network Ablation: This section investigates two elements of the SFCNet: the start stage of the cross-scale feature map guide and the dimension of the fusing feature (see Sec. III-B) for discriminating sperms in the SpermSeg and SpermTrack datasets. The configurations of training and testing follow Sec. IV-A. The results of the ablation study are presented in Tab. IV.

As shown in Tab. IV configurations (a-c, e), the later the start stage of the cross-scale guide, the lower the SFCNet performance, decreasing from 2.74% to 16.54% Dice and 0.44% to 1.81% AP_{50} in SpermSeg and SpermTrack datasets, separately. Such results suggest that the relatively high-dimension

TABLE IV

ABLATION STUDY FOR SFCNET ON SPERMSEG SEMANTIC SEGMENTATION AND SPERMTRACK TRACKING

Configuration	Guide start stage	Resize stage	Dice/%	AP$_{50}$/%
(a)	4	1	47.60	90.29
(b)	3	1	56.25	90.80
(c)	2	1	62.88	91.66
(d)	1	5	37.63	90.15
(e)	1	4	49.80	90.54
(f)	1	3	59.59	91.13
(g)	1	2	63.11	91.78
(e)	1	1	**64.14**	**92.10**

TABLE V

ABLATION STUDY FOR MICROSCOPE MAGNIFICATION ON ADDITIONAL DATA USING SFCNET

Magnifying objective lenses	The number of sperms per image	Dice/%	AP$_{50}$/%
20×	14	60.43	92.59
40×	8	64.03	93.09
100×	2	66.93	97.11

features cannot satisfactorily contribute to supervising the model for learning the representation of small objects without the assistance of low-dimension features.

Furthermore, the influence of the dimension used to fuse multi-scale features for SFCNet performance can be viewed in Tab. IV configurations (d-e). Notably, resizing dimension ceases from 1/32 to 1/2 for fusing feature maps, leading to an increase in Dice by 1.03% $\sim$ 26.51% and AP$_{50}$ by 0.32% $\sim$ 1.95%, in SpermSeg and SpermTrack datasets, individually. The results demonstrate that a comparatively high resize dimension can maintain relatively plentiful information about tiny sperm objects.

2) Microscope Magnification: This section further explores microscope magnification's impact, especially focusing on 20×, 40×, and 100× objective lenses. Adhering to the same settings outlined in Sec. II, we extracted an additional twenty images from videos of the same sample captured separately by 20×, 40×, and 100× objective lenses. The gathered data was annotated at both the pixel and box levels in accordance with Sec. II-B. The model tested was the previously trained SFCNet, which had been used in earlier segmentation and tracking tasks (see Sec. IV-B and Sec. IV-C). The trained SFCNet was then employed in testing, maintaining the same settings as those detailed in Sec. IV-A. The experimental results are in Tab. V.

As demonstrated in Tab. V, the segmentation and tracking performance of SFCNet improves from 60.43% Dice and 92.59% AP$_{50}$% to 66.93% Dice and 97.11% AP$_{50}$% in sperm segmentation and tracking tasks, as the objective lens is increased from 20× to 100×. The results suggest that sperm recognition can be enhanced as the occupied area within the image increases. However, as the magnification level of the objective lenses is increased from 20× to 100×, the number of sperms per image decreases from 14 to 2. This study utilizes 20× objective lenses because it provides acceptable sperm recognition results for distinguishing between healthy and unhealthy sperms, and its relatively large field of view allows for the detection of a more significant number of sperms simultaneously.

V. DISCUSSION

Recent advanced frameworks can be utilized as the encoder portion of the suggested SFCNet algorithm. Our research primarily concentrates on the extended modules for identifying small medical objects. Hence, we only selected the commonly used framework, ResNet50, as the encoder part of SFCNet. Furthermore, the SFCNet could operate as a universal structure to integrate with other segmentation modules (*e.g.*, OCRNet, PSPNet, and LRASPP) in a plug-and-play approach.

Additionally, the SFCNet is based on the multi-scale feature correlation design, which is practical for learning features of objects occupying relatively small regions in images. Therefore, SFCNet has significant potential to be applied to the analysis of other small medical objects such as blood cells, retinal vessels, *etc*.

From a real-time application perspective, the SFCNet takes 0.026 seconds to perform segmentation and tracking tasks for an image on the clinical host computer. In other words, it can achieve 38 FPS when analyzing each frame in an online video captured by a camera. Therefore, it is feasible to use SFCNet for real-time analysis of sperm if the host computer of a clinical device can be similar to or better than ours. We recommend that the host computer integrates an RTX 3090 GPU, a reliable CPU, and other compactable hardware.

VI. CONCLUSION

In this paper, we introduce a novel tiny object recognition network, the SFCNet, to improve the performance of sperm segmentation and tracking. Experimental results suggest that the SFCNet can effectively differentiate sperms and quantitatively measure their morphology and motility parameters. The proposed SFCNet delivered higher performance than other SOTA methods by over 4.19% in Dice and 1.32% in AP. Moreover, the SFCNet achieved errors of less than 15% in analyzing sperm morphometric and motility characteristics, with the exception of a 37.26% error in MAD measurement. Visualization results demonstrate that the SFCNet can accurately detect all sperm locations and distinguish between normal and abnormal sperm. Furthermore, the precise localization and tracking of selected high-quality sperm provide accurate feedback to the automated system for microrobot-assisted reproductive treatment.

REFERENCES

[1] J. B. You, C. McCallum, Y. Wang, J. Riordon, R. Nosrati, and D. Sinton, "Machine learning for sperm selection," *Nature Rev. Urology*, vol. 18, no. 7, pp. 387–403, 2021.

[2] K. Blondeel and P. Houska, *WHO Laboratory Manual for the Examination and Processing of Human Semen*, 6th ed. Geneva, Switzerland: World Health Organization, 2021.

[3] C. Dai et al., "Automated non-invasive measurement of single Sperm's motility and morphology," *IEEE Trans. Med. Imag.*, vol. 37, no. 10, pp. 2257–2265, Oct. 2018.

[4] J. Liu, C. Leung, Z. Lu, and Y. Sun, “Quantitative analysis of locomotive behavior of human sperm head and tail,” *IEEE Trans. Biomed. Eng.*, vol. 60, no. 2, pp. 390–396, Feb. 2013.

[5] V. Chang et al., “Gold-standard and improved framework for sperm head segmentation,” *Comput. Methods Programs Biomed.*, vol. 117, no. 2, pp. 225–237, 2014.

[6] L. F. Urbano, P. Masson, M. VerMilyea, and M. Kam, “Automatic tracking and motility analysis of human sperm in time-lapse images,” *IEEE Trans. Med. Imag.*, vol. 36, no. 3, pp. 792–801, Mar. 2017.

[7] Z. Zhang et al., “Quantitative selection of single human sperm with high DNA integrity for intracytoplasmic sperm injection,” *Fertility Sterility*, vol. 116, no. 5, pp. 1308–1318, 2021.

[8] R. Liu et al., “Interactive dual network with adaptive density map for automatic cell counting,” *IEEE Trans. Autom. Sci. Eng.*, vol. 1, no. 1, pp. 1–13, Nov. 2023.

[9] R. Liu, W. Dai, T. Wu, M. Wang, S. Wan, and J. Liu, “AIMIC: Deep learning for microscopic image classification,” *Comput. Methods Programs Biomed.*, vol. 226, Nov. 2022, Art. no. 107162.

[10] G. Shan et al., “Robotic cell manipulation for blastocyst biopsy,” in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 7923–7929.

[11] H. Liu et al., “Automated morphological grading of human blastocysts from multi-focus images,” *IEEE Trans. Autom. Sci. Eng.*, early access, Apr. 11, 2004, doi: 10.1109/TASE.2023.3264556.

[12] D.-P. Fan et al., “PraNet: Parallel reverse attention network for polyp segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 263–273.

[13] T. Zhou et al., “Cross-level feature aggregation network for polyp segmentation,” *Pattern Recognit.*, vol. 140, Aug. 2023, Art. no. 109555.

[14] W. Dai, R. Liu, T. Wu, M. Wang, J. Yin, and J. Liu, “Deeply supervised skin lesions diagnosis with stage and branch attention,” *IEEE J. Biomed. Health Informat.*, vol. 28, no. 2, pp. 719–729, Feb. 2024.

[15] H. O. Ilhan, I. O. Sigirci, G. Serbes, and N. Aydin, “A fully automated hybrid human sperm detection and classification system based on mobile-net and the performance comparison with conventional methods,” *Med. Biol. Eng. Comput.*, vol. 58, pp. 1047–1068, May 2020.

[16] R. Marín and V. Chang, “Impact of transfer learning for human sperm segmentation using deep learning,” *Comput. Biol. Med.*, vol. 136, Sep. 2021, Art. no. 104687.

[17] C. Dai, G. Shan, H. Liu, C. Ru, and Y. Sun, “Robotic manipulation of sperm as a deformable linear object,” *IEEE Trans. Robot.*, vol. 38, no. 5, pp. 2799–2811, Oct. 2022.

[18] G. Liu et al., “Fast noninvasive morphometric characterization of free human sperms using deep learning,” *Microsc. Microanalysis*, vol. 28, no. 5, pp. 1767–1779, Oct. 2022.

[19] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[20] Z. Zhou, M. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A nested U-net architecture for medical image segmentation,” in *Proc. Int. Workshop Deep Learn. Med. Image Anal.*, vol. 11045, 2018, pp. 3–11.

[21] Y. Yuan, X. Chen, and J. Wang, “Object-contextual representations for semantic segmentation,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 173–190.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[23] J. Wang et al., “Deep high-resolution representation learning for visual recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[24] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder–decoder with atrous separable convolution for semantic image segmentation,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[25] A. Howard et al., “Searching for MobileNetV3,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[27] Z. Tian, C. Shen, H. Chen, and T. He, “FCOS: Fully convolutional one-stage object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.

[28] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[29] Y. Li, Y. Chen, N. Wang, and Z.-X. Zhang, “Scale-aware trident networks for object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6054–6063.

[30] W. Dai, T. Wu, R. Liu, M. Wang, J. Yin, and J. Liu, “Any region can be perceived equally and effectively on rotation pretext task using full rotation and weighted-region mixture,” *Neural Netw.*, vol. 176, Aug. 2024, Art. no. 106350.

[31] K. Wada. (2011). *Labelme: Image Polygonal Annotation With Python*. [Online]. Available: https://github.com/wkentaro/labelme

[32] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[33] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, *arXiv:1706.05587*.

[34] F. Ghasemian, S. A. Mirroshandel, S. Monji-Azad, M. Azarnia, and Z. Zahiri, “An efficient method for automatic morphological abnormality detection from human sperm images,” *Comput. Methods Programs Biomed.*, vol. 122, no. 3, pp. 409–420, Dec. 2015.

[35] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.

[36] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[37] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, “YOLOX: Exceeding YOLO series in 2021,” 2021, *arXiv:2107.08430*.

[38] Y. Bar-Shalom, F. Daum, and J. Huang, “The probabilistic data association filter,” *IEEE Control Syst. Mag.*, vol. 29, no. 6, pp. 82–100, Dec. 2009.

[39] R. Girshick, “Fast R-CNN,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[40] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, “Enhanced-alignment measure for binary foreground map evaluation,” in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 698–704.

[41] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, *arXiv:1412.6980*.

[42] I. Loshchilov and F. Hutter, “SGDR: Stochastic gradient descent with warm restarts,” in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–16.

[43] A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[44] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. (2019). *Detectron2*. [Online]. Available: https://github.com/facebookresearch/detectron2

[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.