Automated Non-invasive Analysis of Motile Sperms Using Cross-scale Guidance Network

Wei Dai[1,2], Zixuan Wu[1,2], Jiaqi Wang[3], Rui Liu[1,2], Min Wang[1,2], Tianyi Wu[1,2], Junxian Zhou[1,2], Zhuoran Zhang[3] and Jun Liu[1,2,*]

***Abstract*— Unbiased measurement of sperm morphometric and motility parameters is essential for assessing fertility potential and guiding visual feedback for microrobotic manipulation. Automated analysis of multiple sperms and selection of an optimal sperm is crucial for in vitro fertilisation treatment, such as robotic intracytoplasmic sperm injection. However, conventional image processing methods have limitations in analysing small sperm objects under microscopic imaging. The emergence of convolutional neural networks (CNNs) has offered promising advancements in sperm analysis. However, previous CNN methods have struggled to accurately segment tiny objects, requiring 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 segmentation network named the cross-scale guidance (CSG) network for accurate and efficient segmentation of minute sperm objects. The CSG network employs innovative modules, including collateral multi-scale convolution, cross-scale feature map guide, and multi-scale feature fusion, to preserve essential sperm details despite their small size. Experimental results indicate that the CSG network surpassed the state-of-the-art models designed for small object segmentation, achieving a significant increase up to 18.62% higher mean intersection over union (mIoU). Additionally, the CSG network excelled in sperm morphometric analysis, achieving errors below 20%. Moreover, sperm motility parameters were further derived from the segmentation results for comprehensive sperm fertility analysis.**

***Keywords*— Automation at micro/nano scale, microrobotics, deep learning, sperm analysis, in vitro fertilisation**

I. Introduction

Infertility is a global health concern that affects millions of couples worldwide. Male factors alone contribute to 30% of fertility cases [1]. The morphology and motility of sperm are critical characteristics in determining its fertility potential and selecting healthy sperm for human reproduction in clinics.

Accurately quantifying sperm morphology and motility is essential for the assessment of sperm quality and treatment of male infertility. The World Health Organisation (WHO) has recommended key morphometric and motility parameters for assessing human sperm, including head area, head length, head width, head ellipticity, tail length, VSL, VCL, VAP, ALH, MAD, LIN, WOB, and STR [2], which are summarised in Fig. 1a. Traditionally, high-magnification microscopy (100× objective) has been used for the subcellular analysis of these morphology parameters [3]. However, under high magnification, the small field of view limits the analysis to one sperm at a time. To obtain unbiased assessment of sperm morphology and motility within 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.

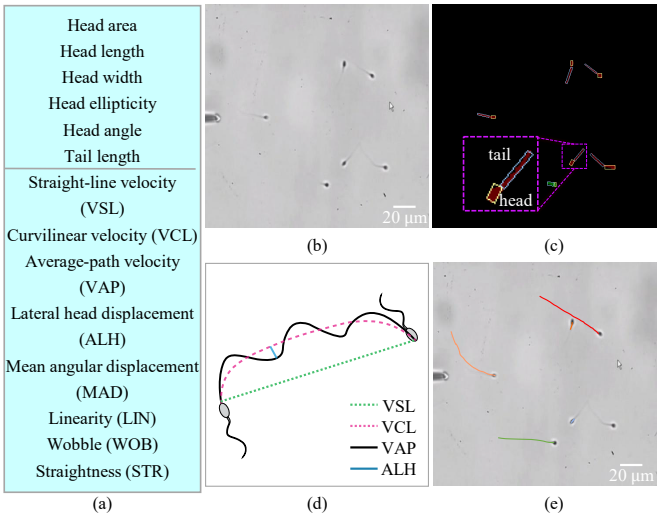


Fig. 1: (a) The quantified parameters representing sperm morphology and motility. (b) An exemplary semen image captured at 20× objective magnification and (c) the accompanying automated morphology analysis using computer vision algorithms. (d) The calculation of sperm motility parameters. (e) An example of detected sperm trajectory.

Recent developments have focused on achieving precise localisation of the sperm head centre using the Kalman filter for quantitative analysis of locomotive behaviour [4]. Additionally, Chang *et al*. [5] employed a k-means algorithm to detect colour variations in culture dishes for identifying sperm heads. The watershed algorithm has also proven valuable in segmenting sperm from the surrounding medium in images [6]. Despite their utility, these methods are ineffective for measuring accurate parameters of sperm morphological structures and motility.

In addition to pixel-based image processing, deep learning algorithms were developed to recognise medical objects [7], [8]. Various deep learning methodologies, such as MobileNet [9], UNet [10], and CNNNet [11], have been utilised to analyse sperm characteristics. However, a common limitation of

*Corresponding author, email: jun.liu@cityu.edu.hk.
[1]Centre for Robotics and Automation, City University of Hong Kong, Hong Kong, China. [2]Department of Mechanical Engineering, City University of Hong Kong, Hong Kong, China. [3]School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), Shenzhen, China.

these methodologies is the requirement for fluorescent tags or staining dyes to enhance sperm visualisation. Unfortunately, using foreign fluorochromes or dyes inevitably damages the cell health, making the sperm clinically impractical.

While Dai *et al*. successfully employed the UNet algorithm to accurately track individual sperm tail for robotic immobilisation [12], and Liu *et al*. used the UNet-tiny model for non-invasive characterisation of sperm head parameters [13], these approaches were limited to analysing either individual sperm head [12] or tail [13] per instance. The non-invasive simultaneous measurement of both morphology and motility parameters for motile spermatozoa has remained largely unexplored.

In this study, we developed a novel deep learning architecture called the cross-scale guidance (CSG) network to differentiate and characterise multiple sperms at 20× objective magnification. The CSG network incorporates four core techniques: collateral multi-scale convolution, cross-scale feature map guide, plug-and-play segmentation module, and multi-scale feature fusion. Importantly, this methodology enabled morphological (Fig. 1c) and motility (Fig. 1e) analysis without the need for fluorescence or dye staining to enhance sperm visibility. Experimental results demonstrate the superior performance of the CSG network, which achieved a mean intersection over union (mIoU) of 51.89 and errors of less than $<20\%$ across all measured morphology parameters. Moreover, the motility parameters calculated based on the segmentation results are further applied to locate healthy sperms.

II. System Setup and Data Acquisition

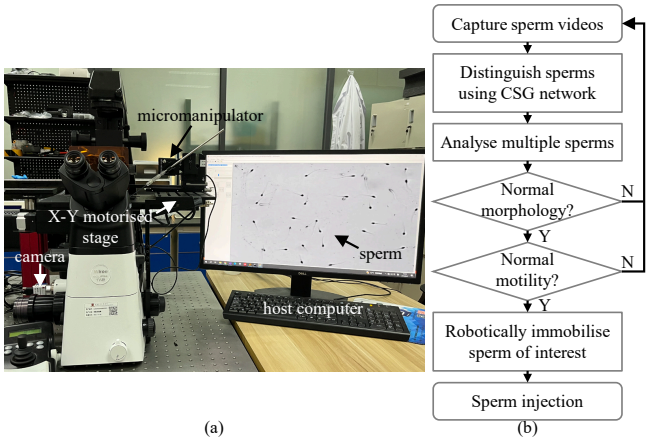
This section presents the configuration of the microrobotic system in Sec. II-A and methods for annotating and processing the data, as expounded upon in Sec. II-B.

A. System setup

The system setup for the sperm analysis and manipulation was built on a standard inverted microscope (Nikon Eclipse Ti2), as depicted in Fig. 2a. A 20× objective lens (Nikon S Plan Fluor, NA: 0.45) was used to achieve microscopic imaging. A CMOS camera (Basler A601f, with a dimension of 640×480) was used for capturing videos at 30 s frames per second for analysis and visual feedback. A motorised 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 immobilisation and injection were conducted with a 3-DOF micromanipulator (MP-285, Sutter Instrument Company) with a positioning resolution of $0.2\,\mu\mathrm{m}$ and a travel range of $25\,\mathrm{mm}$ for each axis.

B. Data Collection and Annotation

In this study, semen samples were obtained from ten volunteers at the Prince of Wales Hospital in Hong Kong. The consent form of the subjects under ethical protocols was obtained. The specimen images were extracted from the captured video clips at a sampling rate of one image



(a) (b)

Fig. 2: (a) System setup for specimen collection and automated sperm infertility analysis. (b) Operation progress of obtaining the sperm of interest.

for 15 frames. Following the WHO guidelines [2], the morphological abnormalities of sperm were categorised into four types: head defects, neck and midpiece defects, tail defects, and excess residual cytoplasm. The ground truth of the sperm entities in the images was carefully annotated by experienced fertility doctors.

In this study, we built a dataset consisting of 148 images with three semantic classes: background, normal sperm (normal), and abnormal sperm (abnormal). The labelled dataset includes 618 instances of normal sperms (42%) and 852 instances of abnormal sperms (58%), totalling 1470 sperms instances. Additionally, due to their small sizes, the sperm cells cover only 1% of the entire image area, leaving the non-sperm background to occupy approximately 99% of the image.

III. Methodology

This section describes the key methodologies for automated sperm analysis with machine learning. The formulation and details of the cross-scale guidance (CSG) model are explained in Sec. III-A.

A. Overall Deep Learning Framework

As shown in Fig. 3, the CSG network consists of four fundamental components: collateral multi-scale convolution (Sec. III-B), cross-scale feature map guide (Sec. III-C), plug-and-play segmentation and muti-scale feature fusion (Sec. III-D). In the initial stage, the input sperm image is processed by a stem module (depicted as the green block in Fig. 3). This module uses a 3×3 convolution operation with a stride of 2 to downsample the original image to a 1/2 tensor shape. To enhance feature extraction for small objects like sperms, the cross-scale feature map guidance module is introduced after the stem module. The CSG network could function as a universal architecture to combine with other segmentation modules (*e.g*., OCR or DeepLabV3) in a plug-and-play manner. In the end, the fusion of the multi-scale features combines all levels of sperm characteristics to provide an accurate segmentation result.

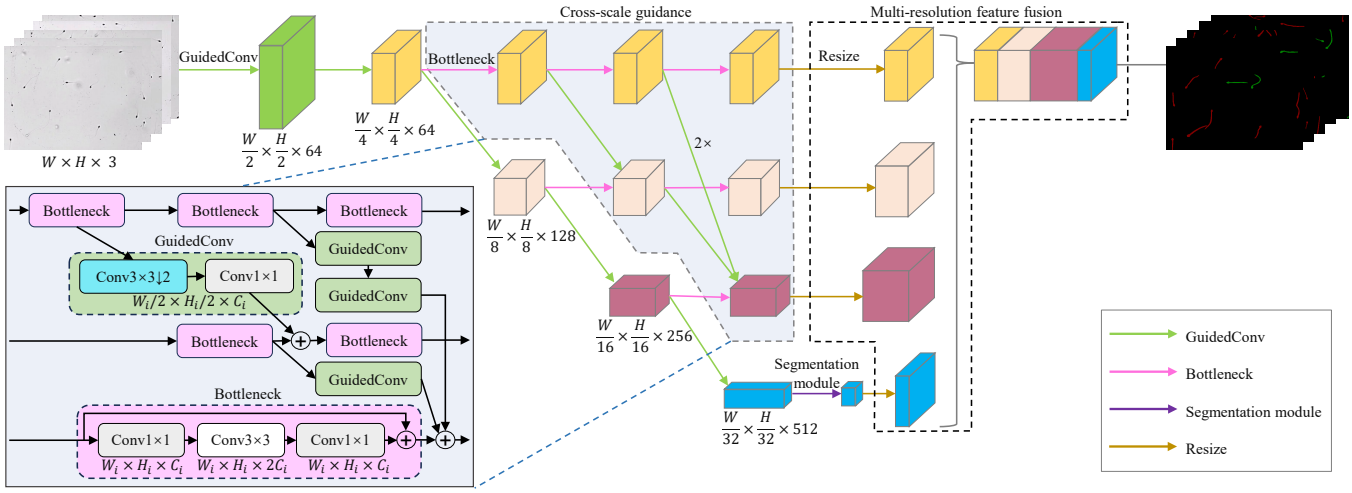


Fig. 3: Architecture of Cross-scale Guidance (CSG) network. The aggregation of low-level features with high-level features through strided convolution is detailly illustrated in the bottom-left region.

B. Collateral Multi-scale Convolution

Computational efficacy is essential in the training and inference phases for neural networks. Therefore, multi-scale convolutions are strategically positioned after the steam module (green block in Sec. III-C). The architecture of this network has four horizontal branches and four vertical stages. Accordingly, the logical organisation can be presented as follows:

$$\begin{array}{ccccccc} B_{11} & \rightarrow & B_{12} & \rightarrow & B_{13} & \rightarrow & B_{14} \\ & \searrow & B_{22} & \rightarrow & B_{23} & \rightarrow & B_{24} \\ & & & \searrow & B_{33} & \rightarrow & B_{34} \\ & & & & & \searrow & B_{44} \end{array} \tag{1}$$

where B_{ij} represents the ith sub-branch and jth cross stage. The output dimensions of the ith branch are exactly $\frac{1}{2^{i+1}}$ of the dimensions of the original input images.

After the image size is reduced to 1/2 of the original dimensions, the processed tensors undergo sequential convolutions, progressing from left to right and top to bottom. This process progressively aggregates features spanning from lower to higher levels in a parallel manner. Consequently, the feature maps encapsulate information from the preceding stages to generate a comprehensive representation.

Both the stem module and the sub-branches B_{11}, B_{22}, B_{33}, and B_{44} function as the encoder component of the proposed segmentation model. Notably, a variety of advanced backbones can be employed in the diagonal branches. In this study, the ResNet50 architecture [14] was selected as the encoder.

C. Cross-scale Feature Map Guide

The inherent challenge for detecting small objects lies in preserving the feature of diminutive entities such as sperms during a sequence of downsampling convolutions using a stride of 2. Hence, it is imperative to harness the potential of features characterised by larger dimensions in the initial stages. This study applies the cross-scale feature maps to the guidance of subsequent stages for representation learning.

An illustration of the feature map guidance spanning three distinct scales (branches) is highlighted in the middle dashed box and light-blue region in Fig. 3. Each bottleneck block (pink arrow in Fig. 3) has operations including one 1×1 convolution followed by a 3×3 convolution and 1×1 convolution with skip connection. Moreover, the Guided Convolution (GuidedConv) block for using the upper-level branch to guide the lower-level branch consists of a 3×3 stride convolution (with a stride of 2) followed by a 1×1 convolution. Within this context, given three input tensors, $\{R_b, b \in \{1, 2, 3\}\}$, the output tensor, R', is calculated through the following equation:

$$R' = f_{1r}(R_1) + f_{2r}(R_2) + f_{3r}(R_3) \tag{2}$$

where the transformation $f_{xr}(R_b)$ performs $(r - x)$ 3×3 convolution operations with a stride of 2.

Because the final stage 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 B_{44} unless explicitly stated.

D. Plug-and-play Segmentation and Feature Fusion

Since of the final branch B_{44} extracts the highest-level visual features of sperm entities, an integral segmentation module is appended at the end of B_{44}. Furthermore, to seamlessly integrate the advanced segmentation algorithms within the CSG network, the segmentation module is designed as a plug-and-play component (indicated by the purple arrow in Fig. 3). In this study, the experimental evaluation included three state-of-the-art segmentation module architectures, including a semantic segmentation system equipped with atrous convolution and conditional random field (DeepLabV3) [15], Object Contextual Representations (OCR) [16], and Lite Reduced Atrous Spatial Pyramid Pooling (LR-ASPP) [17].

The outputs generated at the $r = 4$ stage from B_{14}, B_{24}, B_{34}, and B_{44} have different feature scales. Therefore, a crucial step is to re-sample these outputs to ensure uniform height and width dimensions. Since the output originating

from B_{14} has the dimension most similar to that of the original image, all sub-stream outputs are reshaped to align with the dimensions of the B_{14} output by using the linear interpolation technique. This process is visually depicted by the right dashed box in Fig. 3.

Owing to the relatively low resolution (96 DPI (dots per inch)) of the image acquisition with the 20× objective, the intricate morphology of the small sperm poses a challenge in accurate recognition. In light of these limitations, the analysis is focused solely on specific 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 [18].

Since the occupied region of a sperm head is significantly larger than its tail, the analysis of sperm motility is achieved by tracking the movement of the head. The centre of the bounding box around the sperm head is assigned as the sperm's position. The sperm motility parameters are computed based on the sperm's trajectory (*e.g.* VSL, the velocity along the straight-line path). Because spermatozoa may cross over one another, their trajectories can become interpolated. To ensure correct trajectory mapping of the target sperm, the joint probabilistic data association filter (JPDAF) [19] was applied to associate trajectory points belonging to the same sperm.

IV. Experiment Results and Discussion

In the experiments, the proposed CSG architecture was evaluated and compared with the state-of-the-art machine learning algorithms. The evaluation of segmentation performance employs two primary metrics: intersection over union (IoU) and the mean intersection over union (mIoU). IoU quantifies the overlap between the prediction results and ground truth at the pixel level, calculated by the formula:

$$\text{IoU} = \text{TP}/(\text{TP} + \text{FP} + \text{FN}) \tag{3}$$

where TP, FP and FN represent the true positive, false positive and false negative regions, respectively.

A. Implementation Details

The experimentation for the proposed method was conducted utilising the collected sperm object dataset comprising 148 images with three semantic classes: background, normal sperm (normal), and abnormal sperm (abnormal). These semantic classes were annotated at the pixel level. To facilitate the evaluation, the dataset was divided into two parts, the training and testing sets, with a partition ratio of 4:1 (118 vs. 30 images).

Additionally, the mini-batch size was set to 4. The random crop was employed to resize the input images to a dimension of 512×512. The optimisation process hinged on the Adam optimiser [20], coupled with the adoption of cross-entropy loss. The learning rate was adjusted utilising a cosine schedule [21], decreasing from 5×10^{-5} to 1×10^{-6}. The comprehensive training was executed for 100 epochs, with results computed by averaging three separate training and testing cycles. The experimental computations were performed on an RTX3090 GPU paired with an Intel Xeon Platinum 8375C CPU.

TABLE I: SEGMENTATION IOU AND MIOU (UNIT: %) FOR VARIOUS METHODS.

Method	Module	Background	Normal	Abnormal	mIoU
SegNet	-	99.11	7.11	24.93	42.41
UNet	-	99.30	13.82	34.06	48.25
UNet++	-	99.29	18.29	33.29	48.40
ResNet50	OCR	98.86	0.00	0.00	33.98
	LR-ASPP	98.87	1.36	5.14	35.19
	DeepLabV3	98.86	0.00	0.02	33.27
CSG Network (ours)	OCR	**99.31**	21.23	32.77	51.45
	LR-ASPP	99.30	22.41	33.36	51.64
	DeepLabV3	**99.31**	**21.61**	**34.60**	**51.89**

B. Segmentation Results and Analysis

To evaluate the efficacy of the proposed method, six state-of-the-art (SOTA) tiny-object segmentation models were included as reference points in the experiments. These models were stratified into two categories: (1) backbones coupled with DeepLabV3 [15], OCR [16], and LR-ASPP [17]; (2) encoder-decoder architectures with "U-shape" structure, including SegNet [22], UNet [23], and UNet++ [24].

As illustrated in Tab. I, the CSG network in tandem with DeepLabV3, yielded the highest mIoU score of 51.89, outperforming the SOTA small object segmentation methods. Among all SOTA methods, SegNet, UNet, and UNet++ show an improved IoU performance of up to 15.13% compared to the integration of ResNet50 with DeepLabV3, OCR, and LR-ASPP. This phenomenon underscores the efficacy of the U-shape structure that capitalises on low-level features to enhance the ability of models to recognise tiny objects.

Furthermore, ResNet50 with the tested segmentation modules obtained less than 6% IoU in distinguishing normal or abnormal sperm classes. Although SegNet, UNet, and UNet++ acquired 24.93% $\sim$ 34.06% IoU in segmenting abnormal sperms, these methods are limited in identifying normal sperms, garnering IoU values in the range of 7.11% $\sim$ 18.29%. Conversely, the CSG network with tested segmentation modules achieved IoU values exceeding 21% and 32% for normal and abnormal sperm class segmentation, respectively. The variance in IoU across different sperm types could potentially stem from the relatively fewer normal sperms than the abnormal ones (618 vs. 852).

The CSG network with DeepLabV3 or OCR also achieved superior background segmentation performance, with an IoU of 99.31%. Notably, all other tested methods achieved background segmentation IoU exceeding 95%. This profound difference in IoU between background (>95% IoU) and sperm (<40% IoU) is primarily due to the background covering nearly 99% of the total image area, while the sperm region constitutes less than 2% of the image.

C. Visualisation

In addition to quantitative evaluation, the segmentation results of SegNet, UNet, UNet++, ResNet50 + DeepLabV3,

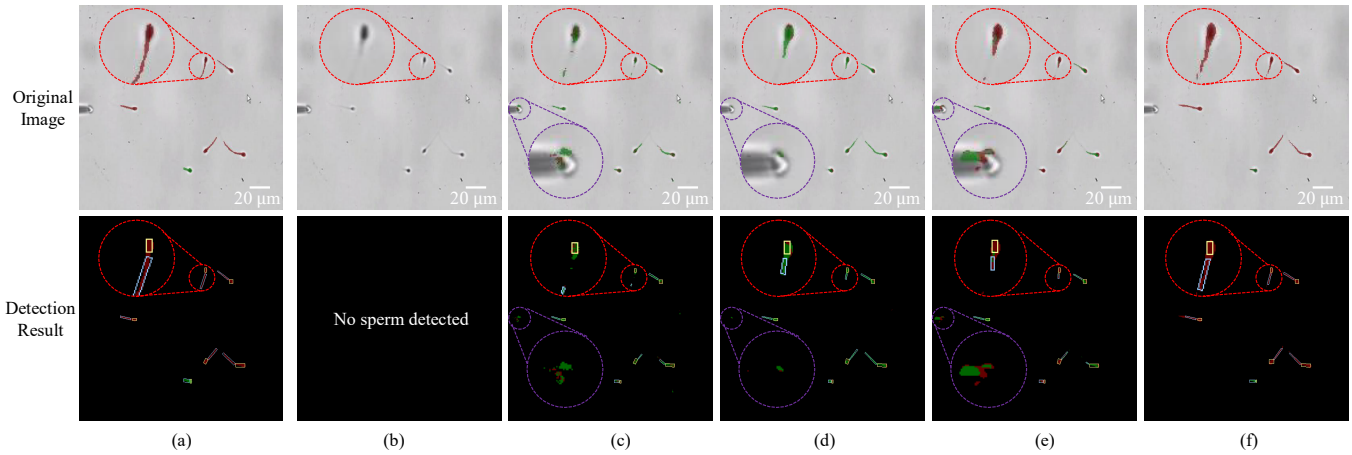


Fig. 4: Visualisation of segmentation ground truth (a) and results using (b) ResNet50 + DeepLabV3, (c) SegNet, (d) UNet, (e) UNet++, and (f) CSG Network + DeepLabV3.

and CSG Network + DeepLabV3 were also compared with an in-depth analysis of sperm morphology.

The prediction results of segmentation masks are exemplified in Fig. 4. It is clear from Fig. 4b that ResNet50 + DeepLabV3 struggles to accurately recognise sperm locations, as evident from images without apparent masking. Conversely, although the SegNet, UNet, and UNet++ algorithms successfully detected all sperm positions in the image (Fig. 4c), they fell short in precisely identifying the heads and tails of the sperm. Notably, SegNet, UNet, and UNet++ misidentified the micropipette as a sperm in the second sample (refer to the third and fourth rows in Fig. 4c-e). In contrast, the CSG network + DeepLabV3 exhibited the capability to identify all sperm positions accurately and effectively reconstruct the morphologies of sperms entities within the image (see Fig. 4f). Moreover, the proposed CSG network + DeepLabV3 classified the micropipette as background.

Furthermore, the sample sperm image has two classes, normal and abnormal, represented by red and green regions in Fig. 4. Although SegNet, UNet, and UNet++ successfully detected all sperm positions, they mistakenly categorised normal sperms as abnormal ones, as visually highlighted in Fig. 4c-e. However, the CSG network + DeepLabV3 proficiently differentiated normal and abnormal sperms, aligning closely with the ground truth, as evidenced by the red and green regions in Fig. 4af.

D. Sperm Morphometric Analysis

To assess the performance of automated segmentation algorithms in the medical application of tested models, the morphometric parameters were measured. The ground truth values of sperms in testing images (321 sperms, 30 images) were measured using ImageJ by averaging annotation from three independent expert technicians. The errors ($\pm$ standard error) associated with automated quantification were assessed across various morphometric parameters using SegNet, Unet, UNet++, ResNet50 + DeepLabV3, and CSG Network + DeepLabV3. These errors are summarised in Fig. 5.

It is evident that ResNet50 + DeepLabV3, in line with the findings in Tab. I and Fig. 4b, exhibited errors exceeding

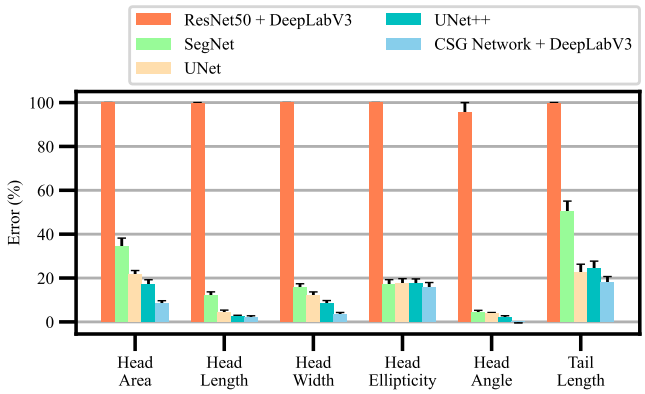


Fig. 5: Errors in automated morphometric analysis using deep learning methods compared to manual benchmark.

95.40% across all parameters, suggesting that the ResNet50 + DeepLabV3 struggled to recognise sperms in the images. However, the proposed CSG Network + DeepLabV3 attained the smallest errors in all sperm morphometric parameters, with percentages ranging from 8.60$\pm$1.06% for head area, 2.22$\pm$0.57% for head length, 3.28$\pm$1.02% for head width, 15.66$\pm$2.29% for head ellipticity, 0.01$\pm$0.00% for head angle, and 17.93$\pm$2.74% for tail length, outperforming the second-best model, UNet++, by a reduction of up to -8.72%.

E. Case Study

To investigate the novel methods in the analysis of sperm structures, a case study was performed by randomly selecting five individual sperm samples. Subsequently, the predictive values of morphological parameters computed by the CSG Network + DeepLabV3 are presented in Tab. II. The calculated morphological parameter values for healthy sample sperms were 14.50$\sim$25.31 µm^2 for head area, 4.75$\sim$5.74 µm for head length, 3.00$\sim$3.75 µm for head length, 1.27$\sim$1.67 AU for head ellipticity, and 33.03$\sim$40.45 µm for tail length. Notably, the most noticeable disparities between prediction results and ground truth were observed in head area and tail length, attributing to the challenges associated with unclear boundaries of sperm objects in low-resolution images. Moreover, the first sample had head parameter values similar to

TABLE II: AUTOMATED QUANTIFICATION OF FIVE SPERM SAMPLES (AU: ARBITRARY UNIT).

Sperm No.	Morphology							Motility									Healthy
	area (μm^2)	length (μm)	Head width (μm)	ellipticity (AU)	angle (°)	Tail length (μm)	Normal	VSL (μm/s)	VCL (μm/s)	VAP (μm/s)	ALH (μm/s)	MAD (°)	LIN (AU)	WOB (AU)	STR (AU)	Normal	
1	14.50	5.62	3.00	1.88	90.00	15.74	✗	12.64	12.64	12.65	0.66	0.62	1.00	1.00	1.00	✓	✗
2	14.50	5.00	3.00	1.67	0.00	38.93	✓	0.68	0.80	0.98	0.82	0.97	0.85	1.22	0.70	✗	✗
3	17.50	5.74	3.50	1.64	2.73	40.45	✓	11.97	12.01	12.01	0.99	0.22	1.00	1.00	1.00	✓	✓
4	19.56	6.25	2.00	3.13	72.25	33.83	✗	0.20	0.20	0.21	0.29	0.47	0.98	1.01	0.98	✗	✗
5	25.31	4.75	3.75	1.27	0.00	33.03	✓	0.76	0.76	0.76	0.10	3.19	1.00	1.00	1.00	✗	✗

healthy spermatozoa, but its tail length fell below 20 μm, one of the characteristics of abnormal sperms.

Furthermore, the motility measurement of the sperm in the right part of Tab. II indicates that sperms No. 2, 4, and 5 hardly moved, with VCL lower than 1 μm/s. Thus, sperms No. 2, 4, and 5 are regarded as abnormal in terms of motility. In contrast, sperms No. 1 and 3 exhibited VCL and VAP over 12 μm/s, and 1.00 for LIN, WOB, and STR, which are within the normal range for sperm characteristics. Although sperm No. 1 was normal in motility, it had abnormal morphology. Thus, sperm No.3 was the only healthy sperm among the five samples.

V. Conclusions

In this paper, we introduce a novel tiny object segmentation network, the CSG network, to enhance the performance of segmenting sperm in medical applications. The experimental results indicate that the CSG network is capable of differentiating sperms by measuring both morphology and motility parameters with high accuracy and efficiency. The proposed CSG network outperformed SOTA methods by over 3.59% mIoU and delivered over 16.45% better mIoU than the conventional ResNet50-based segmentation network. Additionally, the CSG network achieved errors of less than 20% in analysing sperm morphometric characteristics. Visualisation results demonstrate that the CSG network could accurately detect all sperm locations and effectively discriminate between normal and abnormal sperms. Furthermore, the localisation and tracking of selected high-quality sperm offer accurate feedback to the microrobotic system for advanced reproductive treatment.

References

[1] J. B. You, C. McCallum, Y. Wang, J. Riordon, R. Nosrati, and D. Sinton, "Machine learning for sperm selection," *Nature Reviews 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 (sixth edition)*. World Health Organization, 2021.

[3] C. S. Dai, Z. Zhang, J. Huang, X. Wang, W. Meng, J. Zhang, S. Moskovtsev, C. Librach, K. Jarvi, and Y. Sun, "Automated non-invasive measurement of sperm motility and morphology parameters," in *IEEE International Conference on Robotics and Automation*. IEEE, 2018, pp. 2682–2687.

[4] J. Liu, C. Leung, Z. Lu, and Y. Sun, "Quantitative analysis of locomotive behavior of human sperm head and tail," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 2, pp. 390–396, 2012.

[5] V. Chang, J. M. Saavedra, V. Castañeda, L. Sarabia, N. Hitschfeld, and S. Härtel, "Gold-standard and improved framework for sperm head segmentation," *Computer Methods and Programs in Biomedicine*, 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 Transactions on Medical Imaging*, vol. 36, no. 3, pp. 792–801, 2016.

[7] H. Liu, D. Li, C. Dai, G. Shan, Z. Zhang, S. Zhuang, C.-W. Lee, A. Wong, C. Yue, Z. Huang *et al.*, "Automated morphological grading of human blastocysts from multi-focus images," *IEEE Transactions on Automation Science and Engineering*, pp. 1–9, 2023.

[8] W. Dai, R. Liu, T. Wu, M. Wang, J. Yin, and J. Liu, "Deeply supervised skin lesions diagnosis with stage and branch attention," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, 2023.

[9] 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," *Medical & Biological Engineering & Computing*, vol. 58, pp. 1047–1068, 2020.

[10] R. Marín and V. Chang, "Impact of transfer learning for human sperm segmentation using deep learning," *Computers in Biology and Medicine*, vol. 136, p. 104687, 2021.

[11] G. Shan, Z. Zhang, C. Dai, H. Liu, X. Wang, W. Dou, and Y. Sun, "Robotic cell manipulation for blastocyst biopsy," in *International Conference on Robotics and Automation*. IEEE, 2022, pp. 7923–7929.

[12] C. Dai, G. Shan, H. Liu, C. Ru, and Y. Sun, "Robotic manipulation of sperm as a deformable linear object," *IEEE Transactions on Robotics*, vol. 38, no. 5, pp. 2799–2811, 2022.

[13] G. Liu, H. Shi, H. Zhang, Y. Zhou, Y. Sun, W. Li, X. Huang, Y. Jiang, Y. Fang, and G. Yang, "Fast noninvasive morphometric characterization of free human sperms using deep learning," *Microscopy and Microanalysis*, vol. 28, no. 5, pp. 1767–1779, 2022.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

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

[16] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 173–190.

[17] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for MobileNetV3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1314–1324.

[18] 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," *Computer Methods and Programs in Biomedicine*, vol. 122, no. 3, pp. 409–420, 2015.

[19] Y. Bar-Shalom, F. Daum, and J. Huang, "The probabilistic data association filter," *IEEE Control Systems Magazine*, vol. 29, no. 6, pp. 82–100, 2009.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2017, pp. 1–16.

[22] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2015, pp. 234–241.

[24] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *8th ML-CDS Workshop on International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2018, pp. 3–11.