Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Ensembling noisy segmentation masks of blurred sperm images

Emilia Lewandowska [a], Daniel Węsierski [a,d], Magdalena Mazur-Milecka [b], Joanna Liss [f,e], Anna Jezierska [c,a,b,*]

[a] Cameras and Algorithms Lab, Gdańsk University of Technology, Poland
[b] Department of Biomedical Engineering, Faculty of Electronics, Telecommunications, and Informatics, Gdańsk University of Technology, Poland
[c] Department of Modelling and Optimization of Dynamical Systems, Systems Research Institute Warsaw, Poland
[d] Multimedia Systems Department, Faculty of Electronics, Telecommunication, and Informatics, Gdańsk University of Technology, Poland
[e] Invicta Research and Development Center, Sopot, Poland
[f] Department of Medical Biology and Genetics, University of Gdańsk, Poland

ARTICLE INFO

Keywords:
Computer-aided sperm analysis
Microscopic images
Full sperm segmentation
Ensembling
Inter-rater agreement
Noisy labels

ABSTRACT

Background: Sperm tail morphology and motility have been demonstrated to be important factors in determining sperm quality for *in vitro* fertilization. However, many existing computer-aided sperm analysis systems leave the sperm tail out of the analysis, as detecting a few tail pixels is challenging. Moreover, some publicly available datasets for classifying morphological defects contain images limited only to the sperm head. This study focuses on the segmentation of full sperm, which consists of the head and tail parts, and appear alone and in groups.
Methods: We re-purpose the Feature Pyramid Network to ensemble an input image with multiple masks from state-of-the-art segmentation algorithms using a scale-specific cross-attention module. We normalize homogeneous backgrounds for improved training. The low field depth of microscopes blurs the images, easily confusing human raters in discerning minuscule sperm from large backgrounds. We thus propose evaluation protocols for scoring segmentation models trained on imbalanced data and noisy ground truth.
Results: The neural ensembling of noisy segmentation masks outperforms all single, state-of-the-art segmentation algorithms in full sperm segmentation. Human raters agree more on the head than tail masks. The algorithms also segment the head better than the tail.
Conclusions: The extensive evaluation of state-of-the-art segmentation algorithms shows that full sperm segmentation is challenging. We release the SegSperm dataset of images from Intracytoplasmic Sperm Injection procedures to spur further progress on full sperm segmentation with noisy and imbalanced ground truth. The dataset is publicly available at https://doi.org/10.34808/6wm7-1159.

1. Introduction

Many deep learning applications require figure-ground segmentation. The performance of segmentation models varies across modalities and acquisition settings. Our study focuses on segmenting full sperm from the background in blurry images using ensembles of deep neural networks. As embryologists seek sperm with desired shape and motion attributes to increase chances of fertilization, sperm assessment can take advantage of the segmentation task. However, modern deep neural networks are still challenged by blurry microscopic images of minuscule sperm, with spatially uneven contrast, despite their significant progress in binary segmentation in the last decade.

The studied application is important for human well-being. Infertility affects up to 15% of reproductive-aged couples worldwide. It may lead to multiple psychological disorders, including stress, sadness, and depression. The stigmatization of infertile couples is common. Some cultures demand that, for a woman to be socially acceptable, she should have at least one biological child [1]. Male infertility is estimated to contribute to more than half of all global childlessness cases [2]. The most common causes of male infertility are the absence or low sperm levels and abnormal sperm morphology and motility.

Some infertility problems can be solved through In-Vitro Fertilization (IVF), such as Intracytoplasmic Sperm Injection (ICSI). This method involves the injection of a single sperm into an oocyte. The very first step of the whole procedure is sperm selection – a key decision affecting the fertilization outcome. It has been shown that sperm abnormalities correlate with embryo development at later stages [3].

* Corresponding author at: Department of Modelling and Optimization of Dynamical Systems, Systems Research Institute Warsaw, Poland.
E-mail address: anna.jezierska@ibspan.waw.pl (A. Jezierska).

https://doi.org/10.1016/j.compbiomed.2023.107520
Received 22 February 2023; Received in revised form 11 August 2023; Accepted 19 September 2023
Available online 22 September 2023

Therefore, the selection of high-quality spermatozoon is crucial. Meanwhile, the success rate of these Assisted Reproductive Technologies (ART) has plateaued at ∼33% per fertilization cycle [4], largely due to suboptimal sperm selection practices [5].

The sperm quality assessment criteria by the World Health Organization (WHO) are morphology, motility, and vitality [6]. An embryologist performs the assessment analysis, which is subjective, inconsistent, non-repeatable, time-consuming, and costly. In contrast, Computer-Assisted Sperm Analysis (CASA) systems can provide objective and fast semen assessment. The treatment costs are lower than manual analysis, which is currently hardly affordable for financially disadvantaged couples. CASA systems aim to systematically quantify shape and movement and perform statistical analysis after counting sperm according to specific selection criteria.

Most CASA systems classify sperm by their deformations [7,8]. The morphology is an important quality parameter, and the WHO precisely defines a plethora of malformations (Fig. 1). Some CASA systems that use computer vision methods focus on motility analysis [9,10]. Other systems use supervised training of shape malformation classifiers end-to-end, bypassing sperm segmentation. This approach has been validated for the sperm head with some success [11]. Nevertheless, increased black-boxing of a decision process counters the expected explainability in healthcare systems [12]. Additionally, it is challenging for human raters to prepare quality labels for the tail part and for the motility and vitality of groups of sperm. Automatic sperm classification still lacks acceptable precision for widespread clinical use [8].

We suggest two main advantages of using segmentation in the CASA systems. Firstly, sperm segmentation can be viewed as an auxiliary task to quality classification. The segmentation masks increase explainability in the system for sperm quality assessment. We argue the system should know what pixels belong to sperm in the image when deciding sperm quality because segmentation and classification tasks are related [13–15]. Secondly, the raters can use segmentation masks to highlight the shape and motion of blurred sperm parts for improved sperm visibility and parameterization of sperm attributes. In effect, less noisy defect labels should translate, in turn, to better-trained sperm quality classifiers [16].

Most works prioritize efforts towards detecting sperm heads to simplify the problem at the expense of accuracy. Namely, the sperm tail also plays a key role in assessing the abnormal morphology [6] and motility [9,10]. Studies have established that the shapes of flagellar beats determine the movement path [17]. Therefore, analyzing sperm tail's beating patterns can provide new information supporting high-quality cell selection. It is therefore important to develop methods that increase the quality of segmentation of all sperm parts [18].

On the other hand, the segmentation of the flagellum is much more challenging than the segmentation of the head due to the microscopic video characteristics. With small width and quick movements, the flagella are hardly visible under blurry and low-contrast imaging conditions, often confusing the human raters. The collection of noise-free ground truth labels of the tail's shape and motion quality seems daunting and could result in a low inter-rater agreement. In particular, the recording conditions such as brightness and contrast variations (Fig. 2a,b), elongated artifacts (Fig. 2c,f), and overlapping objects (Fig. 2d,e) are a major reason why a labeler finds it difficult to identify the subtle features and defects in tail morphology and motility. The tail should be uniform along its length, thinner than the midpiece, approximately 45 μm long, and should not have a sharp bend or a coil [6,19]. Sperm's tail that is short, multiple, broken, bent, irregular in shape, coiled, or with any combination of these attributes is abnormal [6]. Consequently, many existing detection-based solutions exclude sperm flagellum from computer-aided analysis [7,20]. Moreover, some publicly available datasets for the morphological defects classification contain labels limited only to the head part [21–23].

This work aims to investigate the potential of ensembling state-of-the-art segmentation algorithms in figure-ground sperm segmentation on a new dataset of ICSI images. It is well known that ensembling improves results to some extent. This study describes a neural ensembling approach for joint sperm head and tail segmentation. We analyze the performance of modern binary segmentation algorithms with a particular focus on the accuracy of tail segmentation. We argue the correct segmentation of the flagella will support not only the sperm classification based on the flagella's structure (see Fig. 1j,l–t) but also aid in assessing sperm's motility. Our contributions include:

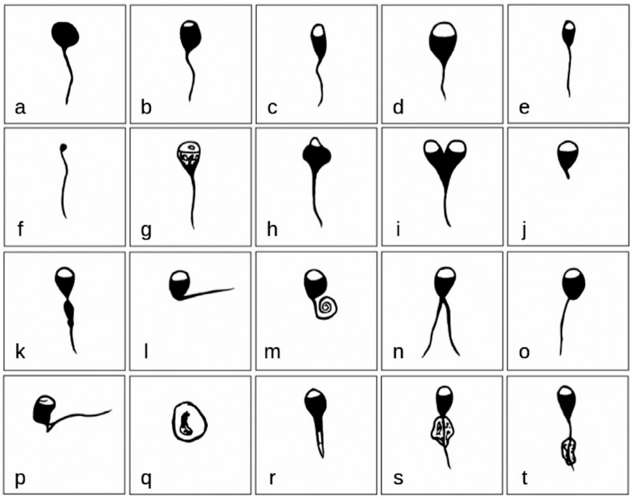

Fig. 1. An atlas of sperm malformations [24,25]: (a) round head/no acrosome, (b) small acrosome, (c) elongated head, (d) megolo head, (e) small head, (f) pinhead, (g) vacuolated head, (h) amorphous head, (i) bicephalic, (j) loose head, (k) amorphous head, (l) broken neck, (m) coiled tail, (n) double tail, (o) abaxial tail attachment, (p) multiple defects, (q) immature germ cell, (r) elongated spermatid, (s) proximal cytoplasmic droplet, (t) distal cytoplasmic droplet.

1. a new full sperm problem formulation and quantitative measure for the evaluation of sperm segmentation results with noisy labels (Section 3);
2. a new ensembling deep neural network, composed of an FPN backbone and a two-branch encoder of the input image and segmentation masks, fused with a new cross-attention module (Section 4) which was shown in an experimental study to outperform the state-of-the-art methods in terms of the quality of the segmentation results (Section 6);
3. an extensive experimental study of deep learning in sperm segmentation problem showing optimal training settings, e.g., image normalization, and limitations of state-of-the-art segmentation methods (Section 6).
4. a new SegSperm dataset of fully labeled sperm in regular ICSI images, with a subset of multiple labels, coupled with ground truth masks aggregation method (Section 5). We make the dataset publicly available at https://doi.org/10.34808/6wm7-1159.

The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 introduces the problem of full sperm segmentation. Section 4 describes the proposed neural ensembling architecture. The dataset is presented in Section 5. The experimental studies are presented in Section 6. In Section 6.3, we describe the results of ensembling segmentation masks, followed by Discussion in Section 6.4. Finally, Section 7 concludes this work.

2. Related work

The problem of sperm segmentation has gained a lot of attention. Traditional computer vision techniques [26–28] and more recently

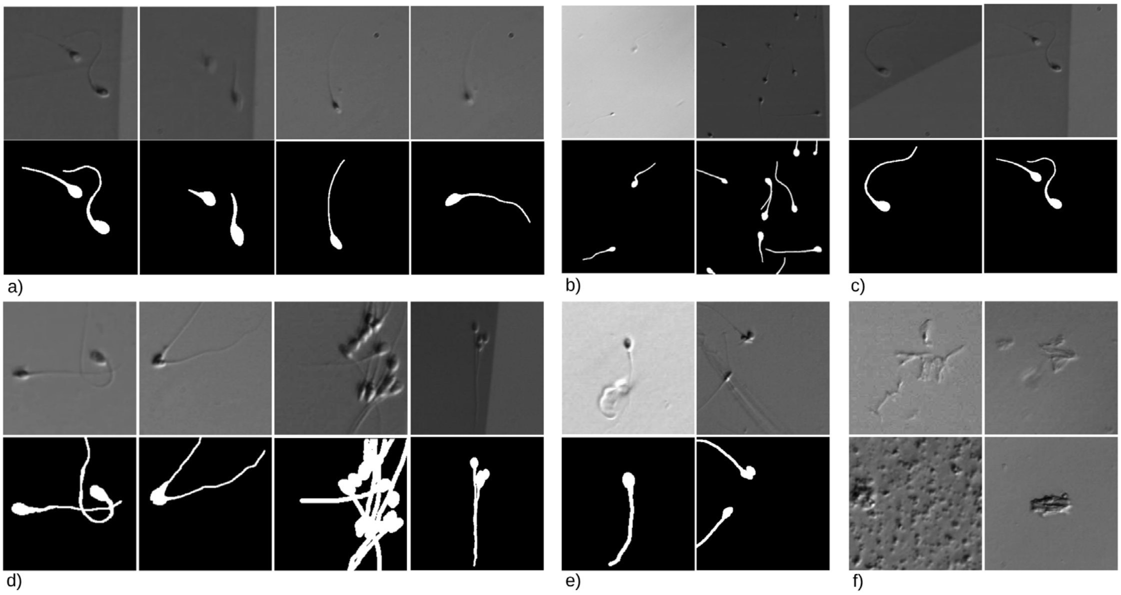

Fig. 2. Different working conditions: (a) blurry semen, (b) differences in brightness, (c) elongated artifacts in the background, such as lines, (d) groups of overlapping sperm, (e) sperm near background artifacts, (f) sperm-like background artifacts.

deep learning techniques [18,29–32] have been proposed. Most sperm segmentation approaches are limited to sperm head [5,22,29,33]. The head segments are used to analyze the motion tracks, parameterized by, e.g., progressive velocity or path linearity, and head morphology analysis [34].

The contributions to the tail segmentation problem have been limited [28,35–39]. The authors proposed two-step, no-training approaches for sperm tail segmentation in early works [35,36]. The first step was to detect the sperm head and its midpoint. The tail was then searched iteratively in the surrounding region. In [36], the first step used the Gaussian Mixtures Model for modeling the appearance of the background and the head and the Bayesian approach for figure-ground segmentation. It was followed by the tail identification step, using a structural similarity index [40] and Rényi entropy [41] in the iterative scheme. Several difficult cases were illustrated in the article, including the detection of tails in images with low contrast. In addition to being highly sensitive to fixed thresholds, modeling sperm as a chain of points compromised the robustness to occlusions and self-occlusions. Video processing techniques were explored in [28]. After background subtraction, the resulting full sperm segments were further refined by morphological filtering. The identified problems included missing tail fragments in the middle or at the end of the tail. The main limitation of this method was its sensitivity to background artifacts with similar color or shape attributes to sperm.

A framework for automatic sperm analysis, grounded on automatic assessment of sperm morphology, including the head, midpiece, and tail parameters, such as tail length, was presented for the first time in [37]. The authors automated the measurement of motility and morphology parameters per single sperm, where motility parameters were computed at low magnifications and morphology parameters at high magnifications. The main challenges were tracking individual sperm when they intersected with each other and segmentation of individual sperm for accurate morphology assessment. The latter problem was approached by a sequential scheme including image restoration, minimizing a quadratic cost function with total variation prior and fuzzy c-means clustering [42]. The limitation of this approach was that the morphology parameters were computed only for target single sperm. Hence, although the morphology of the tail of individual sperm could be assessed, the same analysis could not be performed simultaneously for a group of sperm, a necessity for the sperm selection task.

Recent studies have demonstrated the significance of sperm tail morphology in assessing sperm motility. In [38], the authors found that tail defects correlate with DNA fragmentation levels. In other contributions [43,44], the authors presented a systematic study on the correlation and prediction of sperm DNA integrity from morphological parameters and developed a machine learning framework for predicting DNA fragmentation levels based on sperm morphology. Morphology parameters were confined mainly to standard head parameters. However, these approaches demonstrated the possibility of developing an automatic method for assessing sperm quality without human bias [45] and might serve as a foundation for further development, for example, extending the number of sperm morphological parameters to tail parameters. It is possible to obtain these parameters based on the segmentation of full sperm.

For successful fertilization, the potential benefits of analyzing sperm locomotion [9,46–49] are discussed in [19]. It is concluded that the discovery of sperm locomotion patterns contributes to our understanding of how sperm navigate inside the female reproductive system. These behaviors can be used to design approaches for selecting sperm that are highly fertile. In [49], it is noted that the CASA systems provide only limited insight into sperm motion. Motility is assessed by tracking the movement of the head, ignoring the flagellum. Recent studies have suggested that tail beating may be an effective method of studying the motility of single sperms [19,45]. The pattern analysis of bovine sperm tail motions can be found in [50]. Some novel tail-related parameters have been proposed in [49], such as the frequency of flagellar beats, the speed of flagellar arc waves, and the tail's maximal length and width. Software for tracking sperm using high-speed camera systems has been developed, such as SpermQ [51] or FAST [49]. Since these methods require sperm to be imaged under dark background conditions, analyzing many individual sperm can be challenging [45]. For standard ICSI imaging, however, discovering locomotion patterns involves reconstructing all points along the centerline of the sperm tail. This requires some segmentation of the entire spermatozoon.

Ensembling has been extensively studied in the literature. For the review of the ensembling methods, please see [52], and deep ensembling models for segmentation are discussed in more recent work [53]. As part of our study, we analyze the neural ensembling of multiple variants of inputs, including the input image. A similar formulation was considered in [54]. This formulation produced the optimized final result by concatenating the results of multiple segmentation algorithms

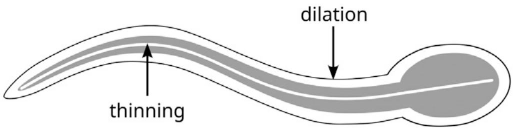

Fig. 3. The thinning of a sperm segmentation mask (gray) yields the white skeleton. The dilation inflates the mask (black contour).

with the input image in the second step. However, the input image is not generally ensembled with segmentation masks in such so-called stacked ensemble models [55–58]. The authors of [57] propose to ensemble the segmentation masks from different model training checkpoints. Instead of using output masks, ensembling models learn from the features extracted by various individual models. The final results will vary based on the features selected for the model. Depending on the number of features in each model, the ensembling may be biased towards some models. Our study uses a more straightforward approach that leverages soft segmentation masks and an image as input. Ensembling has been proposed for sperm quality assessment in [5]. Traditional convolutional neural networks were ensembled for automated classification of human sperm head morphology. We are the first to propose the ensembling of segmentation networks to address the problem of full sperm segmentation.

3. Problem statement

We cast sperm segmentation as a binary classification problem, where the first class is small, elongated objects that are partially out of focus, and the second class is the background. Let $I \in \mathcal{N}^{H\times W}$ and $M \in \{0,1\}^{H\times W}$ denote an input image and a corresponding binary segmentation mask, respectively. The sperm and background pixels are denoted with 1 and 0, respectively, and the pair H, W corresponds to the image height and width. The segmentation $\mathcal{S}$ is a mapping such that $\mathcal{S} : I \mapsto M$.

In the context of deep learning, network $\mathcal{F}_\Theta$ is a mapping $\mathcal{F}_\Theta : I \mapsto S$, where S denotes soft segmentation mask and Θ is a set of parameters. Soft segmentation mask assigns a single-valued score to each pixel in the image within the range of 0 and 1. These heatmaps are loosely interpreted as probabilities of pixels belonging to one of two classes, a foreground object or a background scene. Hence, a common value for the binary classification threshold is 0.5. The following will refer to the binarization mapping as $\mathcal{B} : (S, t) \mapsto M$, where t denotes a threshold.

In the studied microscopic images, the background dominates over the minuscule sperm. The inherent class imbalance thus affects classifier training. The discriminatively trained model miscalibrates the segmentation scores for the minority class [61], as measured e.g. by the stratified Brier score [62], and shifts its optimal classification threshold towards the scores closer to the majority class. We calculate optimal thresholds t for state-of-the-art methods and show that the threshold values vary slightly between methods but are far from the standard $t = 0.5$ threshold.

The segmentation results can be enhanced not only by studying an impact of a threshold but also by reducing the variance of the background. A microscopic image of sperm is dominated by a homogeneous background that can have varying light intensity in ICSI videos. This study considered an image normalization approach to reduce the variance of image intensity and increase contrast helping models focus on sperm shape more than on accounting for the varying light intensity during training.

The segmentation problem defined in Eq. (2) is spatially not equally difficult. Among all pixels corresponding to $M(i,j) = 1$, the pixels presenting sperm head are less blurry than the ones presenting thin, elongated semen tail. Long and flexible sperm tail easily moves out of focus as the microscope has a narrow depth of field [63]. Let M_p and M_r denote binary segmentation masks where the r mask is the reference. We assume the most important pixels of the true segmentation mask lie in the elongated center of the mask. Imprecise labeling at the sperm contours is inevitable as the imaged sperm often has low-contrast and vanishing boundaries. In effect, we develop more optimistic yet practical quality measures for the sperm segmentation problem. To compute recall, the reference mask is thinned (skeletonized) with a thinning operator $t(M)$, and the evaluated p mask is dilated with a $d(M,k)$ operator, where k is a morphological kernel. The reference mask is dilated to compute precision, and the evaluated p mask is eroded with an $e(M,k)$ operator. Fig. 3 illustrates the operations of mask thinning and mask dilation. In the following for the sperm segmentation problem, we propose the asymmetric evaluation measure IoU′ defined as:

$$\begin{aligned} R' &= \text{Recall}(t(M_r), d(M_p,k)) \\ P' &= \text{Precision}(d(M_r,k), e(M_p,k)) \\ \text{IoU}' &= \frac{R' * P'}{R' + P' - R' * P'} \end{aligned} \tag{1}$$

4. Method

We propose a deep neural ensembling network with cross-attention modules for aggregating soft segmentation masks of individual segmentation algorithms. Let $\mathcal{E} : \left(I, \{\hat{S}_i\}_{i=1}^n\right) \mapsto \hat{M}_\mathcal{E}$ denote an ensembling network, where $\hat{M}_\mathcal{E}$ is an estimate of M after ensembling $n \in \mathcal{N}$ different soft segmentation masks $\hat{S}$. The proposed architecture (Fig. 4) re-purposes the Feature Pyramid Network (FPN) backbone [59,64]. The FPN has found various applications, among others, in object detection [64], multi-class segmentation [59], and instance segmentation [65,66]. We propose a new adaptation of FPN by augmenting it with attention modules and applying it to ensemble-based full sperm segmentation.

The proposed FPN-based architecture has two encoder branches. The first branch serves feature extraction from an input image I. The second branch extracts features from a stack of soft segmentation masks $\{\hat{S}_i\}_{i=1}^n$. We use three levels of a pyramid feature. We denote each pyramid feature level with s ranging from 1 to 3. The spatial dimension is reduced by a factor $\times\frac{1}{2}$. Two bottom-up paths are integrated with the same structure and feature dimension.

Additionally, we introduce a cross-attention block (Fig. 5). The main goal of this block is to combine features from both branches by leveraging the channel and spatial information between them. To this end, we leverage the spatial and channel attention module (CBAM) from [60]. Specifically, to point out *what* is meaningful in the input feature, the Channel Attention Module aggregates spatial context descriptors by average-pooling and max-pooling operations. Channel attention is focused on meaningful spatial attention in the area of interest. The introduced channel attention mechanism produces an output added to top-down pathways' features. The Spatial Attention Module involves average-pooling and max-pooling operations in the channel axis space and concatenates them to generate feature descriptors highlighting *where* are the informative regions.

Our method uses three separate attention branches, one per scale. Each branch gets the corresponding attention mask. Each attention mask is multiplied by the input feature maps to obtain the final output separately. After that, at each sale level, we apply two convolution operations. Dimension features are upsampled to the input dimension and concatenated. The outputs of each scale are concatenated as well. Finally, convolution filters were applied to reduce the number of output channels.

For the proposed architecture, the optimal set of parameters Θ is derived by minimizing the loss $\mathcal{L}$:

$$\mathcal{L}(\Theta) = \sum_k^K \sum_i^H \sum_j^W L\left(S^k(i,j), \hat{S}^k_\Theta(i,j)\right), \tag{2}$$

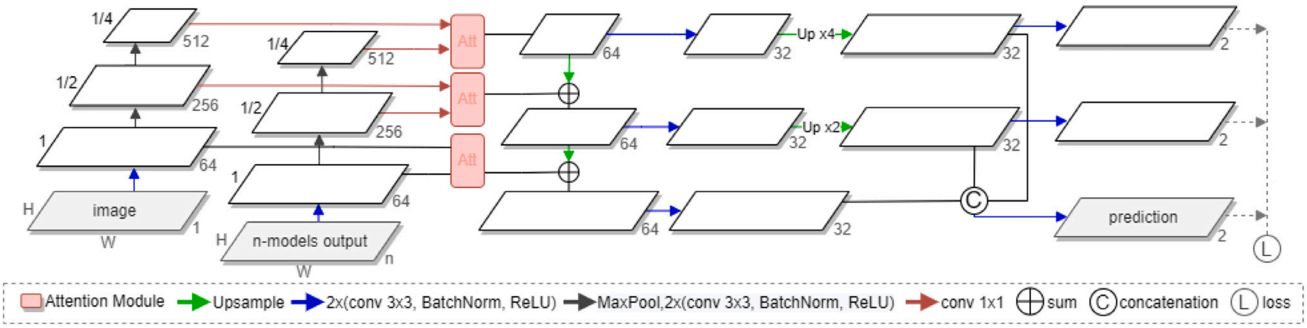

Fig. 4. Architecture of our ensembling deep neural network. The network with FPN backbone [59] has a two-branch encoder structure. The first branch extracts features of an input image with dimensions $H \times W$. The second branch computes features of the concatenated n soft segmentation masks with dimensions $H \times W$, which are output by different segmentation algorithms. At each pyramid level, features from two branches are passed to the cross-attention module (red boxes).

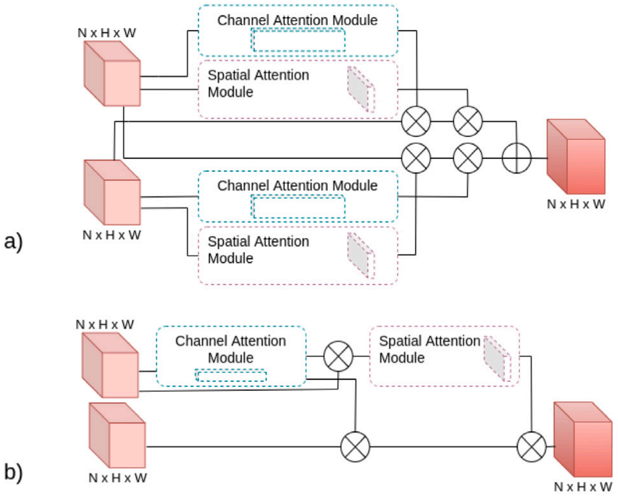

Fig. 5. A cross-attention module is proposed in two variants (V1 and V2). It incorporates features from two network branches. The cross-attention module is based upon the spatial and channel attention modules (CBAM) from [60]. Variant V1 cross-weights both network branches with the associated channel and spatial modules. The red boxes represent input and output feature maps. According to variant V2, the channel and spatial modules refine the image feature branch based on the features from the soft mask segmentation branch.

Table 1
Summary of SegSperm dataset. The dataset consists of gray images with 512 × 512 resolution.

Sets	#Videos	Resolution	#Segmentation masks		
			GT1	GT2	GT3
Train	40	512 × 512	432	0	0
Test	9	512 × 512	119	23	23

Table 2
Inter-rater agreement on 23 sperm segmentation masks. The masks belong to the test set with triple labels from three raters GT1–3. The lower agreement between raters for the strict IoU measure (top) considerably increases for the optimistic IoU′ measure (bottom), computed with kernel size 3 × 3. The IoU′ measure indicates the agreement is high for the head part and moderate for the tail part.

IoU			
	GT1 *vs.* GT2	GT1 *vs.* GT3	GT2 *vs.* GT3
Full	0.5767	0.5790	0.5661
Head	0.7016	0.6635	0.7339
Tail	0.4806	0.4957	0.4702
IoU′ (Eq. (1))			
Full			
	GT1	GT2	GT3
GT1	1	0.8356	0.8541
GT2	0.8428	1	0.8304
GT3	0.8485	0.8307	1
Head			
GT1	1	0.9838	0.9866
GT2	0.9155	1	0.9638
GT3	0.8802	0.9376	1
Tail			
GT1	1	0.7684	0.7888
GT2	0.7856	1	0.7624
GT3	0.7961	0.7648	1

where $\hat{S}^k_\Theta \in [0,1]^{H\times W}$ is an estimate of ground truth S^k for kth sample in training dataset D and distance measure $L(a,\hat{a})$ given by:

$$L(a,\hat{a}) = -\sum_{s=1}^{3} \alpha_s \left(a\log(\hat{a}_s) + (1-a)\log(1-\hat{a}_s)\right) \qquad (3)$$

where $\hat{a}_s$ is an up-sampled estimate of a at scale s and an impact of each scale is weighted by a factor $\alpha_s \in (0,\infty)$.

5. Dataset

Microscopic images of sperm were acquired by Invicta.[1] The original videos recorded the whole ICSI injection procedure. We chose video frames from intervals of the sperm selection phase. The frames were selected based on sharpness degree, large spermatozoa appearance, and background variability with artifacts like lines, spills, and stains.

The SegSperm dataset consists of 551 gray images with binary ground truth masks of sperm. The training set consists of 432 images from 40 videos, and the test set consists of 119 images from 9 videos. The binary masks of spermatozoa were segmented manually by one rater GT1. In addition, 23 images of sperm from the validation set were annotated using the same annotation tool by two more raters, GT2 and GT3. Table 1 summarizes the dataset.

The intersection-over-union measure IoU∈ $[0,1]$, which compares binary segmentation masks by precision and recall, informs about the moderate agreement between three raters in Table 2. The head part is easier to detect than the tail part, as evidenced by the average difference in IoU= 0.2175 between the parts, while IoU is below 0.6 for the full sperm across all three rater-to-rater comparisons. Consequently, the certainty of the ground truth masks of GT1 may raise initial concerns about the protocol that evaluates and ranks algorithms for the sperm segmentation task. The visual inspection of the 23 human segmentation masks of GT1, GT2, and GT3 reveals that human raters disagree mostly at the contours of sperm and the tail end. The raters

[1] www.invictaclinics.com.

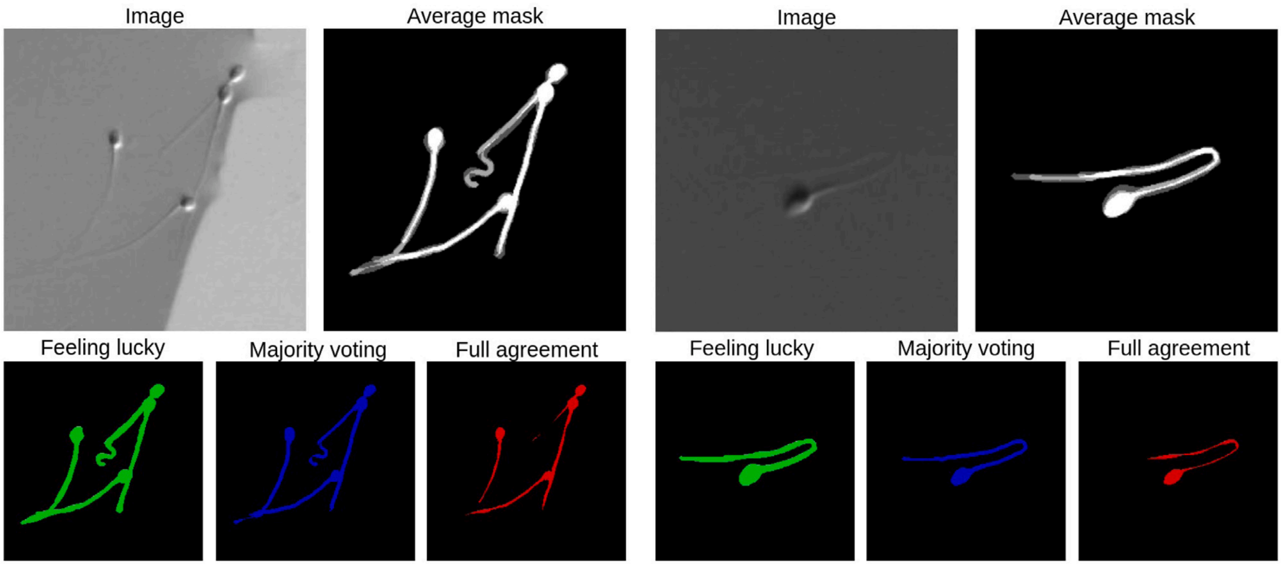

Fig. 6. Three variants of ground truth segmentation masks after aggregating masks from three raters GT1–3 with the following protocols: feeling lucky (at least one rater is positive), majority voting (majority is positive), and full agreement (all raters are positive). Feeling lucky is the most optimistic, majority voting balances between recall and precision, and full agreement is the most conservative. The ground truth masks of three mask aggregation variants differ, especially at the end of the semen tail part, and to some extent impact the quantitative comparison between segmentation methods.

rarely confuse debris with sperm. The agreement between the human raters considerably increases according to the measure IoU′ compared to the standard, strict IoU measure, as shown in Table 2 for the kernel k of size 3 × 3. As the IoU′ depends on the kernel k, it increases further with a larger kernel size. The mean IoU′ scores over six pairs of rater-to-rater comparisons at the kernel size of 3 × 3 are $0.8404, 0.9446, 0.7777$ for full sperm, and head and tail part, respectively, while they increase to $0.9315, 0.9847, 0.8554$ at the kernel size of 5 × 5.

Thus increased inter-rater agreement suggests that the raters consistently agree in segmenting most parts of sperm. They mostly disagree on the sperm contours, where labeling errors stem partly from manual imprecision rather than clinical misjudgment. The tail part, though, which is a marker of sperm motility, can be more troublesome. The raters disagree mostly on the vanishing ending of the tail. Therefore, apart from evaluating the segmentation algorithm on ground truth masks of rater GT1, we propose three ground truth aggregation variants of masks GT1–3, as shown in Fig. 6. In particular, *Feeling lucky* aggregates segmentation masks such that at least one rater has to be positive. It is the most optimistic label aggregation heuristic, with more false positives than the other two variants. *Majority voting* is the common label aggregation method, where the majority is positive. It balances true and false positives. *Full agreement* requires all raters to agree at a pixel. It is the most conservative ground truth aggregation variant, with the lowest number of false positives at the cost of missed true positives.

6. Experiments, evaluation, and results

This section describes experiments that evaluated the potential of ensembling segmentation algorithms for figure-ground segmentation of sperm in microscopic images. In our experiments, we selected 10 state-of-the-art figure-ground segmentation methods with publicly available code: Unet [67], FC-DenseNet [68,69], Unet++ [70], AttUNet [71], Multi-scale Attention-Net (MAnet) [72], Spatial Attention Unet (SA-Unet) [73], DeepLabV3+ [74], Feature Pyramid Network (FPN) [59,64], Unet3+ [75] and UC-TransNet [76].

The performance of segmentation models is evaluated with the *average precision* (AP) measure, which summarizes the precision and recall of a model irrespective of a given threshold. We thus avoid searching for an optimal threshold during model selection for validation. Additionally, we report optimal intersection-over-union scores IoU_{opt} and $\mathrm{IoU}_{0.5}$ for the binary segmentation masks M that are resulting from binarization of S with optimal $t = t_{opt}$ and $t = 0.5$ thresholds, respectively.

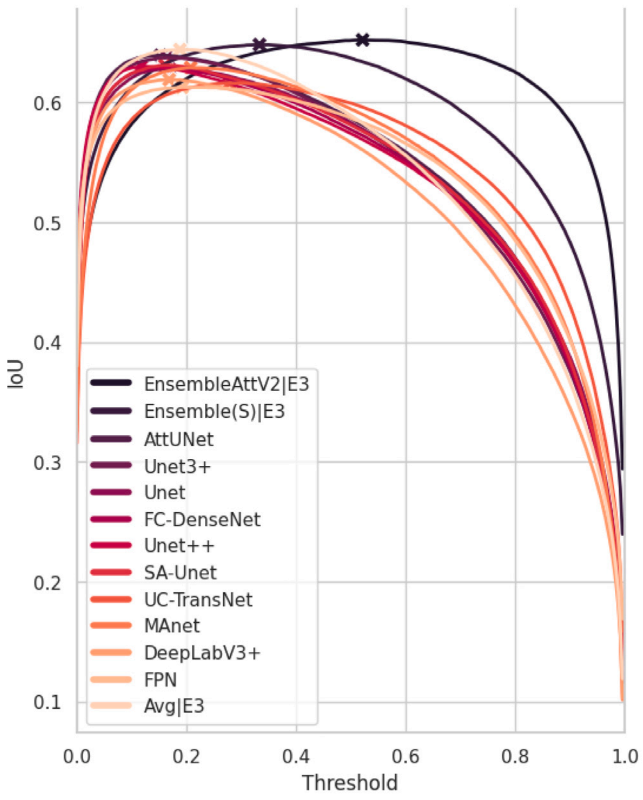

Fig. 7. Dependence of segmentation performance (IoU) on the selected threshold (best viewed in color). Optimal thresholds, which lead to the best IoU score on the test set, are marked with a cross.

The models were trained with the binary cross-entropy loss using the Adam optimizer [77], without pretraining on another dataset. The training used an initial learning rate of 0.001 on a multistep

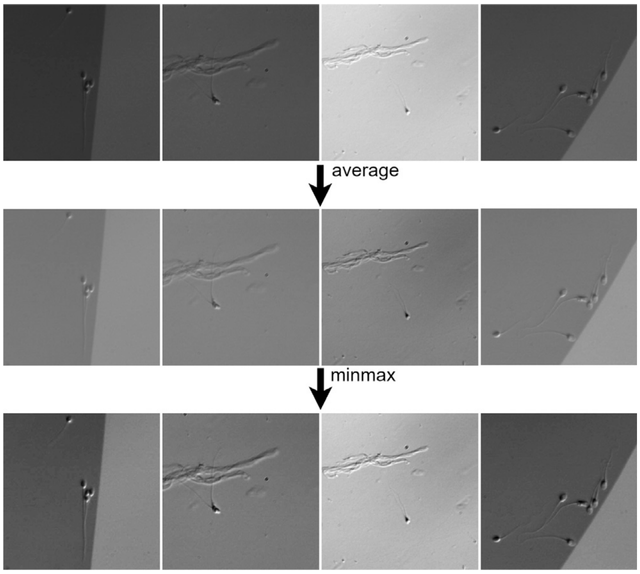

Fig. 8. Image normalization. Original images (top row) are processed to zero-mean images (middle row) that share similar histograms. Minmax increases their contrast (bottom row). For example, the two original images in the middle come from the same video sequence but significantly vary in brightness. After the image-specific minmax normalization, the dataset has a lower variation in brightness and higher contrast.

schedule with milestones 30 and 80 and gamma parameters 0.3. We used batch normalization during training. We set the batch size to 5 and training epochs to 300. The ensemble models expected different training parameters than the state-of-the-art segmentation models. We reduced the learning rate earlier, with milestones at 5 and 15. We also set the batch size to 6 for training the ensemble models. We additionally experimented with other batch sizes (2, 4, 8), a combination of DICE and binary cross-entropy losses, and with the focal loss and class-instance weighting (1:2, 1:3, 1:5) to account for class imbalance. Still, the results were either on par or worse. The size of our ensemble models was 4M parameters, the computational efficiency reached 33 fps. All experiments were run on NVidia GeForce RTX 2080Ti and NVidia Quadro RTX 5000. The models were implemented in PyTorch.

The training of segmentation models for elongated objects [78] can be improved by a careful choice of data augmentation. In the following, we propose to use Gaussian blur (random selection of kernel size in the range of 2–10), zooming in and out (with scale factor ×0.67), rotation, and horizontal and vertical flipping. Each of the transformations was applied with a probability of 50%. We further augmented the training data by contrast and brightness changes up to ±20% and ±30%, respectively. Furthermore, we explored two image normalization variants during training and validation and reported performance improvements in all segmentation algorithms.

The Section is organized as follows. First, we study the optimal training settings, including normalization and threshold in Section 6.1. We also report inter-rater agreement as the apparent variability between the manually segmented masks relates to the uncertainty of evaluation protocols. By adopting tailored evaluation measures and data normalization techniques, our rigorous evaluation showed that the proposed ensembling consistently outperforms individual segmentation algorithms (Section 6.3). Finally, we discuss the results in Section 6.4.

6.1. Optimal thresholds

Multiple factors affect the performance of deep neural networks for object segmentation, ranging from data preprocessing to training to mask postprocessing and evaluation protocols in the presence of noisy labels. We carefully search for improved training of the segmentation methods as segmented masks of sperm are used by and compared to ensembling. Here, we present the impact of the optimal threshold on the final results. Fig. 7 illustrates the optimal thresholds of segmentation models for IoU measure that were trained on our training set and evaluated on the test set (Table 1). The optimal thresholds vary slightly between all segmentation methods and cluster near 0.2, far from the typical threshold of 0.5.

6.2. Image normalization

The ten state-of-the-art models were trained without image normalization and with two image normalization variants: mean subtraction and mean subtraction followed by minmax range extension (Fig. 8). Fig. 9 illustrates the validation results of the architectures that show the impact of image normalization variants on trained models. Training with minmax image normalization led to the highest AP results for all models, often keeping a significant margin of ΔAP ~ 1–1.5 over models trained without image normalization. Consequently, we adopted this normalization procedure for training all methods in the following quantitative and qualitative analyses.

6.3. Ensembling segmentation masks

The most straightforward approach to ensembling is averaging the outputs of multiple classifiers to expect performance improvement. On the other hand, we show that a deep neural network can ensemble soft segmentation masks of individual algorithms further to improve the segmentation accuracy of sperm in microscopic images. We validate our findings by ranking 10 segmentation methods and 24 ensembling variants according to evaluation measures defined in Eq. (1). Each result shows the mean and standard deviation computed on the test set for models from 5 training runs with random initialization. As generalization and robustness are prime deployment aspects, we examine their robustness to progressively blurred images of sperm. The numbers **in bold** indicate the best results in the tables.

The evaluation of segmentation methods in Table 3 ranks AttUnet and Unet3+ as the top two performers according to the AP and IoU_{opt} measures. The next four best methods, Unet, FC-DenseNet, Unet++, and SA-Unet, are worse by ΔAP ~ 0.5–1.0. DeepLabV3+ achieves the lowest AP on our dataset, being inferior to AttUnet by ΔAP ~ 2.5. The AP measure ranks the methods similar to the IoU_{opt} measure that indicates a larger advantage of the attention mechanism of AttUnet over other methods. The reported results for $\mathrm{IoU}_{0.5}$ measure with the segmentation threshold of 0.5 ranks the methods differently, signifying the importance of properly selecting the decision threshold when training on heavily imbalanced datasets.

Soft segmentation masks differ between the methods in Fig. 10. The algorithms react differently on the non-sperm, elongated patterns such as borders between two homogeneous backgrounds and on sperm image regions. The differences are especially visible at the end of tails, the most blurry and the hardest to detect part of sperm. These results suggest that ensembling can potentially reduce the number of false positives, thereby improving the final segmentation masks. While false negatives can be challenging, we note that the ensembling models take soft segmentation masks as input, thus without hard thresholding, allowing for improvement of the final masks by raising the input soft scores.

6.3.1. Selection and fusion of inputs

The selection of soft segmentation masks and the fusion of inputs affect ensembling performance. Four groups of inputs with six fusion configurations evaluate the performance of ensembling. The four arrangements of inputs to the second branch of the network (Section 4) are all masks (E1), single best mask (E2), three best masks (E3), and three best masks on train set (E4). The fusion configurations are the following: averaging soft segmentation masks (Avg, S), ensembling by

Table 3
Comparison of state-of-the-art segmentation models with our ensembling configurations. We specify four sets of segmentation masks E1–4 for ensembling. Two ensembling models with a single encoder branch input a set of segmentation masks (S) and a concatenation of an input image with a set of segmentation masks (I+S). The other two ensembling models (I,S), with the first encoder branch for the input image and the second encoder branch for the set of segmentation masks, fuse features between the branches at each scale either by concatenation (concat) or our two cross-attention modules (AttV1 and AttV2).

Full semen segmentation results of base models and their ensembling configurations E1–4

Base model	AP	$IoU_{0.5}$	IoU_{opt}	E1	E2	E3	E4
AttUnet [71]	**0.8334** ± 0.0023	0.5876 ± 0.0043	**0.6348** ± 0.0043	✓	✓	✓	
Unet3+ [75]	0.8327 ± 0.0036	0.5739 ± 0.0122	0.6286 ± 0.0137	✓		✓	✓
Unet [67]	0.8273 ± 0.0056	0.5710 ± 0.0141	0.6268 ± 0.0034	✓		✓	
FC-DenseNet [68]	0.8258 ± 0.0036	0.5811 ± 0.0035	0.6219 ± 0.0042	✓			
Unet++ [70]	0.8240 ± 0.0021	0.5795 ± 0.0072	0.6278 ± 0.0012	✓			✓
SA-Unet [73]	0.8222 ± 0.0035	0.5688 ± 0.0069	0.6224 ± 0.0051	✓			
UCTransNet [76]	0.8191 ± 0.0067	0.5932 ± 0.0076	0.6255 ± 0.0070	✓			✓
MAnet [72]	0.8138 ± 0.0065	0.5936 ± 0.0075	0.6194 ± 0.0062	✓			
FPN [59]	0.8126 ± 0.0025	0.5851 ± 0.0059	0.6146 ± 0.0033	✓			
DeepLabV3+ [74]	0.8073 ± 0.0054	0.5837 ± 0.0097	0.6103 ± 0.0052	✓			

AP results (mean+stddev) of five models with ensembles of four sets of segmentation masks E1–4

Ensemble model	Inputs	AP (E1)	AP (E2)	AP (E3)	AP (E4)
Avg	S	**0.8397**	**0.8368**	0.8344	0.8303
Ensemble	S	0.8340 ± 0.0025	0.8350 ± 0.0012	**0.8444** ± 0.0008	**0.8424** ± 0.0007
Ensemble	I+S	0.8318 ± 0.0018	0.8279 ± 0.0007	0.8392 ± 0.0018	0.8377 ± 0.0014
Ensemble+concat	I,S	0.8260 ± 0.0019	0.8310 ± 0.0019	0.8392 ± 0.0015	0.8389 ± 0.0015
EnsembleAttV1	I,S	0.8190 ± 0.0038	0.8284 ± 0.0022	0.8348 ± 0.0022	0.8346 ± 0.0024
EnsembleAttV2	I,S	0.8305 ± 0.0025	0.8332 ± 0.0015	0.8419 ± 0.0022	0.8369 ± 0.0039

IoU results (mean+stddev) of five models with ensembles of four sets of segmentation masks E1–4

Ensemble model	Inputs	$IoU_{0.5}$ (E1)	$IoU_{0.5}$ (E2)	$IoU_{0.5}$ (E3)	$IoU_{0.5}$ (E4)
Avg	S	0.5967	0.5922	0.5932	0.5939
Ensemble	S	0.6353 ± 0.0019	0.6426 ± 0.0035	0.6440 ± 0.0036	0.6442 ± 0.0035
Ensemble	I+S	0.6384 ± 0.0025	0.6395 ± 0.0023	0.6398 ± 0.0031	0.6424 ± 0.0046
Ensemble+concat	I,S	0.6286 ± 0.0054	0.6347 ± 0.0074	0.6386 ± 0.0033	0.6412 ± 0.0064
EnsembleAttV1	I,S	0.6337 ± 0.0050	0.6401 ± 0.0036	0.6406 ± 0.0058	0.6447 ± 0.0030
EnsembleAttV2	I,S	0.6305 ± 0.0060	0.6382 ± 0.0069	0.6413 ± 0.0079	0.6304 ± 0.0076

Ensemble model	Inputs	IoU_{opt} (E1)	IoU_{opt} (E2)	IoU_{opt} (E3)	IoU_{opt} (E4)
Avg	S	**0.6443**	0.6410	0.6442	0.6423
Ensemble	S	0.6382 ± 0.0012	0.6455 ± 0.0015	0.6469 ± 0.0017	0.6472 ± 0.0009
Ensemble	I+S	0.6394 ± 0.0023	0.6436 ± 0.0014	0.6434 ± 0.0010	0.6465 ± 0.0015
Ensemble+concat	I,S	0.6351 ± 0.0020	0.6454 ± 0.0015	0.6448 ± 0.0024	**0.6486** ± 0.0011
EnsembleAttV1	I,S	0.6360 ± 0.0041	0.6429 ± 0.0026	0.6451 ± 0.0022	0.6469 ± 0.0019
EnsembleAttV2	I,S	0.6407 ± 0.0043	**0.6462** ± 0.0004	**0.6475** ± 0.0029	0.6485 ± 0.0008

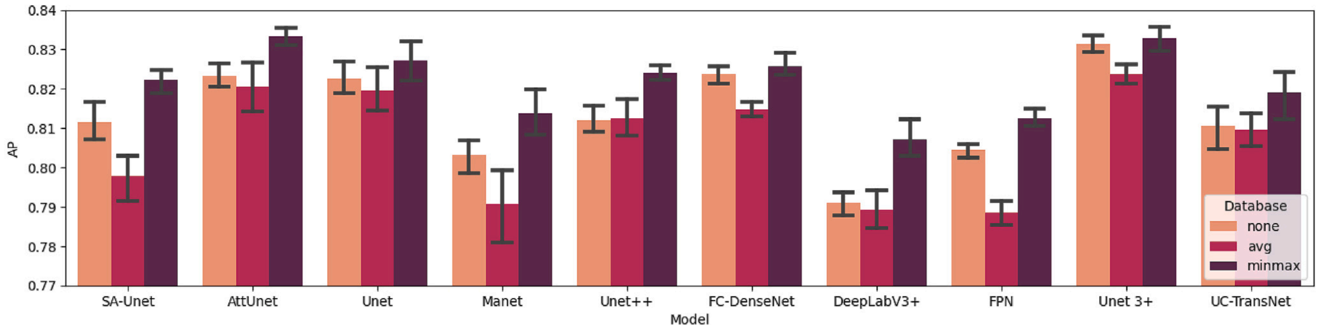

Fig. 9. Performance (AP) comparison of ten segmentation methods, without image normalization and after applying image normalization: mean subtraction (avg) and mean subtraction followed by minmax range extension (minmax).

concatenation of soft segmentation masks in the single-branch network (Ensemble, S), concatenation of soft segmentation masks and input image (Ensemble, I+S), features concatenation from masks and image branches (Ensemble+concat, I,S), combining the features of both branches by cross-attention V1 as shown in Fig. 5a (EnsembleAttV1, I,S) and by cross-attention V2 as shown in Fig. 5b (EnsembleAttV2, I,S).

The results of the experiments are presented in Table 3. As the results are obtained after 5 independent training runs, the masks to all ensembling variants come from a model that achieved the top score out of the 5 runs.

As indicated by AP and IoU_{opt} measures, ensembling by averaging is most effective when input comprises all segmentation masks (E1). Ensembling is least effective with the single, top model (E2). Best AP performance is achieved by ensembling the top three base models (E3). Although the E4 configuration of base models achieves the highest average IoU_{opt} result, the top-performing ensembling from a single trial is attributed to the E3 configuration. The fusion of inputs affects segmentation accuracy as well. In three cases, E2–4, ensembling segmentation masks with an input image works best. Then, the network with the attention module V2 outperforms the attention module V1 in

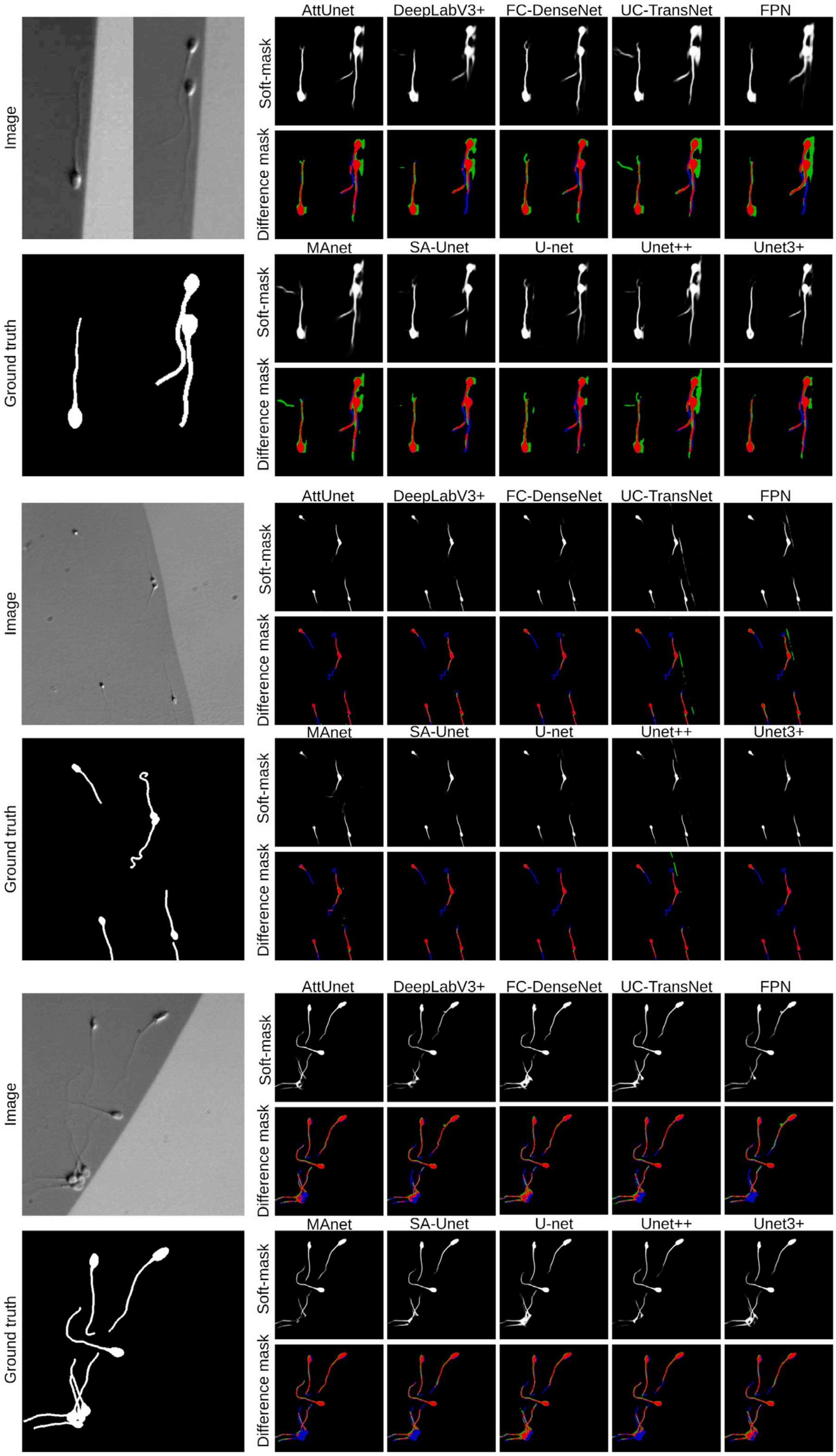

Fig. 10. Sperm segmentation masks of state-of-the-art methods can differ substantially in ICSI images. All methods detect sperm heads mostly correctly. The sperm tail is troublesome. Some methods are sensitive to background artifacts. Difference maps of segmentation results wrt the ground truth labels of annotator GT1: red, blue, and green regions represent true positive (TP), false negative (FN), and false positive (FP) pixels, respectively.

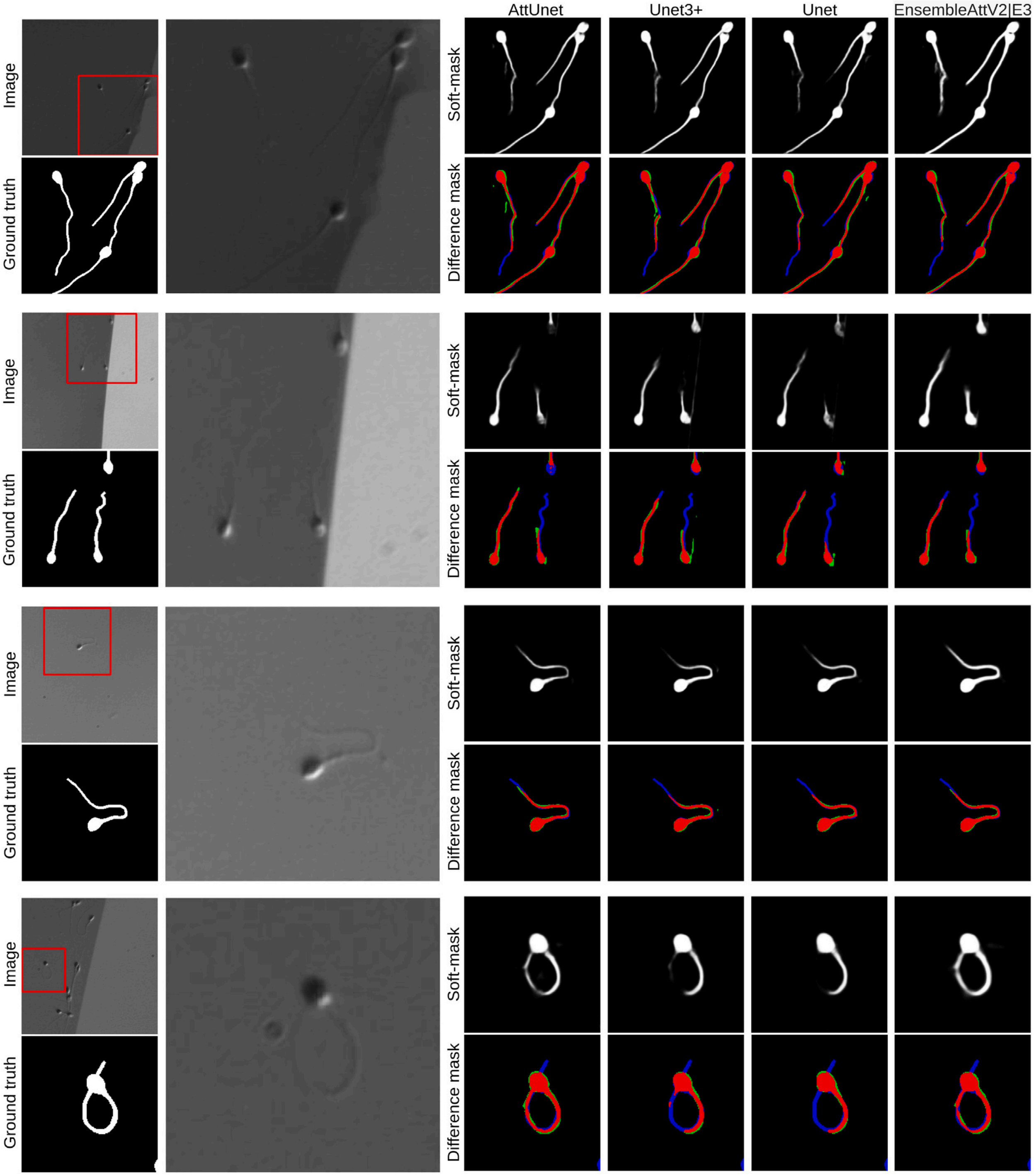

Fig. 11. Comparison between base models of the E3 ensemble (Table 3) and EnsembleAttV2|E3. Ensembling the soft segmentation masks of base models shows that EnsembleAttV2|E3 can select the best segments locally in the presence of false positives (rows 1 and 2) and thickens the masks (rows 3 and 4). Difference maps of segmentation results wrt the ground truth labels of annotator GT1: red, blue, and green regions represent true positive (TP), false negative (FN), and false positive (FP) pixels, respectively.

all cases for both AP and IoU$_{opt}$ measures. Our ensembling model uses the FPN backbone. The FPN algorithm achieves AP = 0.81 and IoU$_{opt}$ = 0.61 and most of our ensembling models achieve AP > 0.83 and IoU$_{opt}$ > 0.63. The qualitative results of the best ensemble model EnsembleAttV2 with E3-selection of base models are presented in Fig. 11. The ensemble model reduces mistakes of the top three base models. It thickens the masks and generates fewer false positives.

6.3.2. Noisy labels of blurred images

The relationship between image sharpness and the quality of predicted segmentation masks of full sperm is shown in Fig. 12. Gaussian kernels with sizes of 1–15 blurred the 119 validation images of rater GT1 from the ICSI dataset. After blurring the images with each kernel, the algorithms segmented the sperm and were compared to the segmentation masks of rater GT1. The ensembling algorithm EnsembleAttV2|E3 consistently maintains its advantage over other algorithms despite the increasing blur in the images until kernel size 9. The results spread after kernel sizes 6, showing that some algorithms are more robust to blur than others. In general, blurry images lower the precision and recall of all segmentation algorithms. Tail segmentation is thus a challenging image segmentation task.

The evaluation protocol must address the ground truth uncertainty that increases with blurry images. To study the ranking consistency of segmentation methods, evaluated on multiple ground truth variants, we quantify the performance of the methods on 23 images from the validation set on (i) individual ground true segmentation masks of raters GT1–3 and (ii) ground truth variants that aggregate masks of GT1–3 in three ways, as shown in Fig. 6. In Table 4, we compare the performance of AttUnet, which achieves top accuracy on our dataset (Table 1), with the ensembling methods Avg|E3 and EnsembleAttV2|E3. Six variants of ground truth segmentation masks evaluate the accuracy of predicted segmentation masks obtained at the optimal binarization threshold. Head segmentation is far easier than tail segmentation for all methods across all ground-truth variants, as in Table 2. Ensembling improves segmentation over the top performing AttUnet in all cases. EnsembleAttV2|E3 is slightly worse than Avg|E3 in half of the cases for the head part but considerably better for the tail part.

6.4. Discussion and limitations

Object segmentation algorithms generally focus on images that contain cluttered backgrounds of indoor and outdoor scenes. The background of ICSI microscopic images is generally homogeneous, with rare local artifacts and non-local light intensity variations. Despite this seemingly simplified setting, image-based sperm segmentation still challenges state-of-the-art deep neural networks, as the appearance of sperm can be non-discriminative. The tail part is especially prone to blur, contrasting little with the background. Precise segmentation of blurred tails is difficult even for human raters (Section 6.3.2).

The improved accuracy of segmentation algorithms by image-specific normalization (Section 6.1) suggests that potential further accuracy gains align with improving the normalization method. Invoking a straight line detector allows image normalization to separate two homogeneous regions of different mean intensities for obtaining more homogeneous images. However, we used state-of-the-art straight line detection [79] on ICSI images that confused lines with elongated artifacts and sperm. Moreover, curved lines in ICSI images separate two homogeneous regions of different intensities, making image normalization more difficult. Then, highly blurred lines bias the estimation of mean intensities of separated regions.

Noisy human labeling and the resulting inter-rater disagreement affect the evaluation protocols that use uncertain ground truth and the supervised learning regimes that penalize incorrect network predictions based on human labels. However, incorporating noisy label training [16] will help little in the presence of false negative errors that raters consistently make. It thus limits the potential of ensembling that prefers more false positives and fewer false negatives. False positives of human raters are rare, though, mostly resulting from lower annotation precision due to fatigue and often indecisiveness due to subjective perception of high-contrast boundaries. Repetitive false negatives across

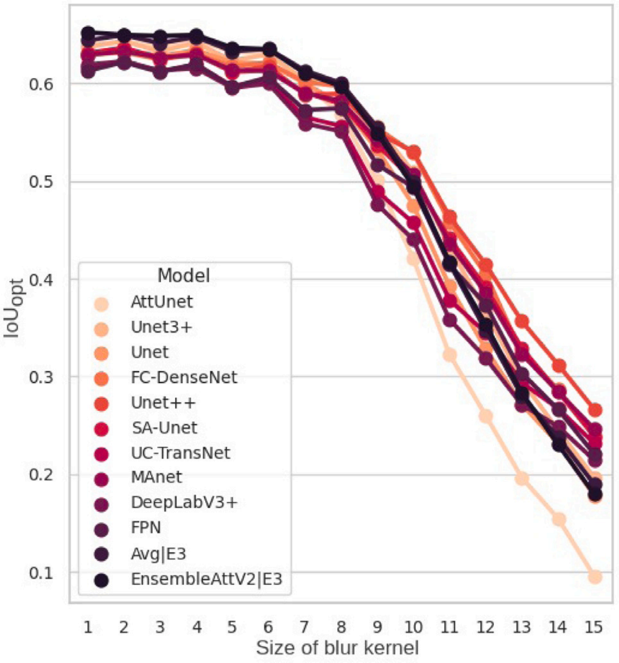

Fig. 12. Assessing the robustness of full sperm segmentation to varying blur levels. Our best ensembling model EnsembleAttV2|E3 consistently outperforms ten state-of-the-art methods across lower magnitudes of an artificial blur.

Table 4
Semen segmentation results (IoU$_{opt}$) on 23 test images with triple labels GT1–3. Comparison between the best base model (AttUnet), ensembling by averaging (Avg|E3), and ensembling by cross-attention module (EnsembleAttV2|E3) wrt masks of raters GT1–3 and wrt three aggregation variants of ground truth masks. For all methods across all evaluation variants, head segmentation is far easier than tail segmentation due to less blur. The results between methods vary more for the tail than for the head part. Generally, our ensembling method EnsembleAttV2|E3 outperforms the other two approaches.

IoU$_{opt}$ head

GT variants	AttUnet	Avg|E3	EnsembleAttV2|E3
GT1	0.7589	0.7649	**0.7875**
GT2	0.7251	0.7352	**0.7422**
GT3	0.7452	**0.7522**	0.7506
Full agreement	0.5724	**0.5784**	0.5612
Majority voting	0.7151	**0.7260**	0.7156
Feeling lucky	0.7821	0.7886	**0.8008**

IoU$_{opt}$ tail

GT variants	AttUnet	Avg|E3	EnsembleAttV2|E3
GT1	0.4850	0.4871	**0.4976**
GT2	0.4354	0.4355	**0.4458**
GT3	0.4939	0.4913	**0.5022**
Full agreement	0.3853	**0.3939**	0.3911
Majority voting	0.5389	0.5437	**0.5491**
Feeling lucky	0.4955	0.4932	**0.5023**

IoU$_{opt}$ full

GT variants	AttUnet	Avg|E3	EnsembleAttV2|E3
GT1	0.6007	0.6054	**0.6210**
GT2	0.5557	0.5609	**0.5723**
GT3	0.6145	0.6176	**0.6270**
Full agreement	0.4930	**0.5030**	0.4957
Majority voting	0.6388	0.6482	**0.6496**
Feeling lucky	0.6165	0.6185	**0.6305**

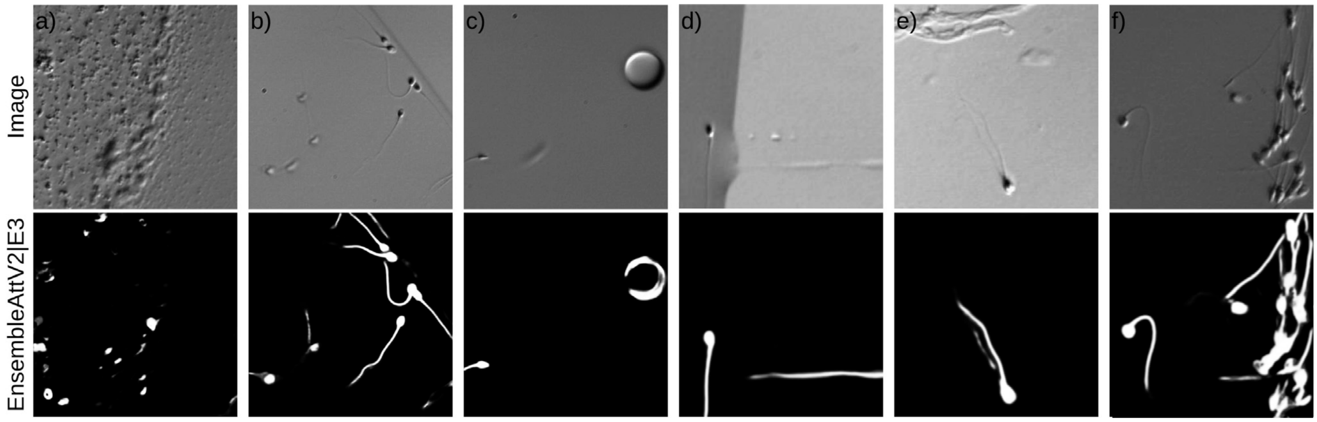

Fig. 13. Soft masks of the ensembling model EnsembleAttV2|E3 on out-of-dataset images. Cluttered backgrounds with head-like patches (a,b), straight lines (b), high-contrasted curves (c), elongated spills (d), and overlapping sperm (e,f) are challenging working conditions that mislead segmentation algorithms. On the positive side, many head-like patches were correctly classified as background (a), borders between dark and light backgrounds were also classified as background despite thin, elongated patterns (b,d), and some stains were correctly classified as background (c,e).

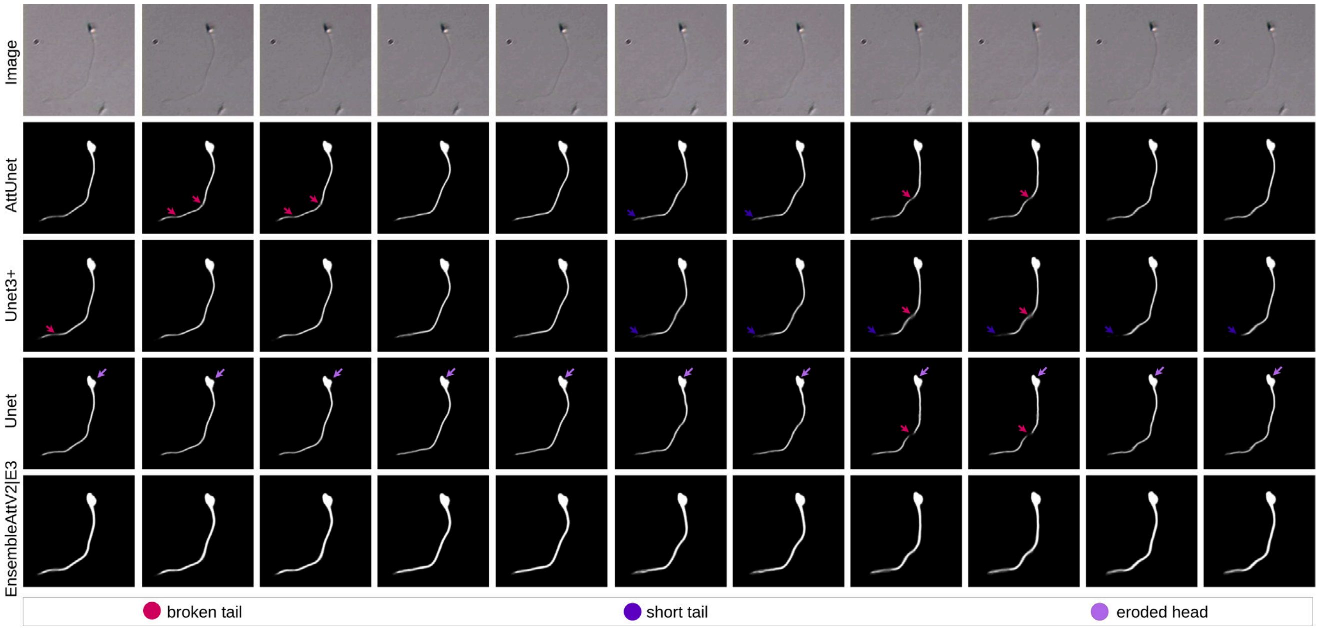

Fig. 14. Sperm segmentation in a video snippet. We define three visible error categories: broken tail, short tail, and eroded head. AttUnet and Unet3+ output masks with brief but frequent temporal inconsistencies. The Unet struggles to maintain temporally consistent segmentation masks, with eroded heads and disappearing tails in the middle and end tail parts. EnsembleAttV2—E3 segments the thickest masks that are most consistent over time.

images are common, particularly for the tail part. This raises questions about whether additional multi-rater labeling of training data and multi-label training regimes add value to the trained models for ICSI image-based sperm segmentation.

Sperm shape is a composition of an egg-like head and a wavy, thin tail. Annotating such tiny image structures requires much manual effort and is time-consuming. Because ICSI image backgrounds are mostly homogeneous, casting the sperm segmentation problem as anomaly detection via one-class classification could avoid laborious image labeling. Moreover, soft segmentation masks of one-class classifiers could add extra diversity to segmentation masks for ensembling. To verify this, we trained a state-of-the-art one-class model [80]. We observed that it could find sperm heads but also artifacts, thereby increasing the false positives. However, the tail part was not discriminative enough for the network. It was neglected, thereby not decreasing the false negatives and having no effect on ensembling the segmentation masks from binary and one-class classifiers. We argue the shape attributes of sperm can be confused quite easily with ICSI image artifacts in one-class classification.

Qualitative analysis in Fig. 13 indicates that air bubbles, elongated spills, and curved high-contrast artifacts can be hard to discern from sperm. Further failure occurs when cells are very close to each other forming tangled tails in large groups. Figure-ground segmentation of crowded regions is challenging, suggesting that datasets from ICSI procedures yield a good testbed for instance-level segmentation under heavy occlusions [81].

Processing videos of ICSI procedures opens new avenues for research. The main ICSI acceptance criterion is sperm motility. Still, sperm or slowly moving sperm is rejected from further procedures. Motion is thus an important cue for (i) filtering out many false positive detections as numerous artifacts are static and (ii) reducing false negatives [82] due to blur. A video-based analysis in Fig. 14 displays temporal inconsistencies of soft mask predictions. Ensembling produces sperm segments that are the most consistent across time, as evidenced by the number of maintained pixels in the soft segmentation masks. However, all image-based segmentation methods struggle to produce error-free segments of a moving sperm cell over time.

7. Conclusions

This study described a deep neural network for ensembling soft segmentation masks. It was successfully applied to the sperm segmentation problem in blurry microscopic images. The ensembling network performed better than state-of-the-art figure-ground segmentation networks on our new dataset. Microscopic imaging blurs the images of sperm, especially at the sperm tail part, which can move out of focus. Blurred object parts can confuse human raters in correct manual segmentation leading to higher inter-rater disagreement. Our future work will go beyond image analysis and focus on segmenting the sperm and computing the sperm motion from videos.

Declaration of competing interest

None Declared

References

[1] W. Cui, Mother or Nothing: the Agony of Infertility, Vol. 88, World Health Organization. Bulletin of the World Health Organization, 2010, p. 881.

[2] M.C. Inhorn, P. Patrizio, Infertility around the globe: new thinking on gender, reproductive technologies and global movements in the 21st century, Hum. Reprod. Update 21 (4) (2015) 411–426.

[3] E. Blahová, J. Máchal, L. Máchal, I. Milaković, Hanuláková, Eliminating the effect of pathomorphologically formed sperm on resulting gravidity using the intracytoplasmic sperm injection method, Exp. Ther. Med. 7 (2014) 1000–1004.

[4] R. Nosrati, P.J. Graham, B. Zhang, J. Riordon, A. Lagunov, T.G. Hannam, C. Escobedo, K. Jarvi, D. Sinton, Microfluidics for sperm analysis and selection, Nat. Rev. Urol. 14 (12) (2017) 707–730.

[5] L. Spencer, J. Fernando, F. Akbaridoust, K. Ackermann, R. Nosrati, Ensembled deep learning for the classification of human sperm head morphology, Adv. Intell. Syst. 4 (10) (2022) 2200111.

[6] T. Cooper, E. Noonan, S. von Eckardstein, J. Auger, H. Baker, H. Behre, T. Haugen, T. Kruger, C. Wang, M. Mbizvo, K. Vogelsong, World health organization reference values for human semen characteristics, Hum Reprod Update 16 (3) (2009) 231–245.

[7] S. Javadi, S. Mirroshandel, A novel deep learning method for automatic assessment of human sperm images, Comput. Biol. Med. 109 (2019) 182–194.

[8] V. Chang, L. Heutte, C. Petitjean, S. Härtel, N. Hitschfeld, Automatic classification of human sperm head morphology, Comput. Biol. Med. 84 (2017) 205–216.

[9] J. Liu, C. Leung, Z. Lu, Y. Sun, Quantitative analysis of locomotive behavior of human sperm head and tail, IEEE Trans. Biomed. Eng. 60 (2) (2012) 390–396.

[10] H.-F. Yang, X. Descombes, S. Prigent, G. Malandain, X. Druart, F. Plouraboué, Head tracking and flagellum tracing for sperm motility analysis, in: International Symposium on Biomedical Imaging, IEEE, 2014, pp. 310–313.

[11] J. Riordon, C. McCallum, D. Sinton, Deep learning for the classification of human sperm, Comput. Biol. Med. 111 (2019) 103342.

[12] J. Amann, A. Blasimme, E. Vayena, D. Frey, V.I. Madai, P. Consortium, Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, BMC Med. Inform. Decis. Mak. 20 (2020) 1–9.

[13] Q. Xu, Y. Zeng, W. Tang, W. Peng, T. Xia, Z. Li, F. Teng, W. Li, J. Guo, Multi-task joint learning model for segmenting and classifying tongue images using a deep neural network, IEEE J. Biomed. Health Inf. 24 (9) (2020) 2481–2489.

[14] Z. Kong, M. He, Q. Luo, X. Huang, P. Wei, Y. Cheng, L. Chen, Y. Liang, Y. Lu, X. Li, et al., Multi-task classification and segmentation for explicable capsule endoscopy diagnostics, Front. Mol. Biosci. 8 (2021) 614277.

[15] Y. Zhou, H. Chen, Y. Li, Q. Liu, X. Xu, S. Wang, P.-T. Yap, D. Shen, Med. Image Anal. 70 (2021) 101918.

[16] D. Karimi, H. Dou, S.K. Warfield, A. Gholipour, Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis, Med. Image Anal. 65 (2020) 101759.

[17] B.M. Friedrich, I.H. Riedel-Kruse, J. Howard, F. Jülicher, High-precision tracking of sperm swimming fine structure provides strong test of resistive force theory, J. Exp. Biol. 213 (8) (2010) 1226–1234.

[18] R. Marín, V. Chang, Impact of transfer learning for human sperm segmentation using deep learning, Comput. Biol. Med. 136 (2021) 104687.

[19] C. Dai, Z. Zhang, G. Shan, L.-T. Chu, Z. Huang, S. Moskovtsev, C. Librach, K. Jarvi, Y. Sun, Advances in sperm analysis: techniques, discoveries and applications, Nat. Rev. Urol. 18 (8) (2021) 447–467.

[20] I. Iqbal, G. Mustafa, J. Ma, Deep learning-based morphological classification of human sperm heads, Diagnostics (Basel) 10 (5) (2020) 325.

[21] F. Ghasemian, S.A. Mirroshandel, S. Monji-Azad, M. Azarnia, Z. Zahiri, An efficient method for automatic morphological abnormality detection from human sperm images, Comput. Methods Programs Biomed. 122 (3) (2015) 409–420.

[22] V. Chang, J.M. Saavedra, V. Castañeda, L. Sarabia, N. Hitschfeld, S. Härtel, Gold-standard and improved framework for sperm head segmentation, Comput. Methods Programs Biomed. 117 (2) (2014) 225–237.

[23] F. Shaker, Human sperm head morphology dataset, Mendeley Data (2017).

[24] A. Sathananthan, Visual Atlas of Human Sperm Structure and Function for Assisted Reproductive Technology, La Trobe and Monash Universities, Singapore: National University, Melbourne, 1996.

[25] G. Palermo, Q. Neri, T. Takeuchi, S. Hong, Z. Rosenwaks, Textbook of Assisted Reproductive Technologies, Informa UK, 2009.

[26] K.S. Park, W.J. Yi, J.S. Paick, Segmentation of sperms using the strategic hough transform, Ann. Biomed. Eng. 25 (2) (1997) 294–302.

[27] H. Carrillo, J. Villarreal, M. Sotaquira, A. Goelkel, R. Gutierrez, A computer aided tool for the assessment of human sperm morphology, in: IEEE International Symposium on BioInformatics and BioEngineering, IEEE, 2007, pp. 1152–1157.

[28] R. Medina-Rodríguez, L. Guzmán-Masías, H. Alatrista-Salas, C. Beltrán-Castañón, Sperm cells segmentation in micrographic images through lambertian reflectance model, in: International Conference on Computer Analysis of Images and Patterns, Springer, 2015, pp. 664–674.

[29] M.S. Nissen, O. Krause, K. Almstrup, S. Kjærulff, T.T. Nielsen, M. Nielsen, Convolutional neural networks for segmentation and object detection of human semen, in: Scandinavian Conference on Image Analysis, Springer, 2017, pp. 397–406.

[30] R.A. Movahed, M. Orooji, A learning-based framework for the automatic segmentation of human sperm head, acrosome and nucleus, in: International Iranian Conference on Biomedical Engineering, IEEE, 2018, pp. 1–6.

[31] R. Melendez, C.B. Castañón, R. Medina-Rodríguez, Sperm cell segmentation in digital micrographs based on convolutional neural networks using U-Net architecture, in: International Symposium on Computer-Based Medical Systems, 2021, pp. 91–96.

[32] Q. Lv, X. Yuan, J. Qian, X. Li, H. Zhang, S. Zhan, An improved U-Net for human sperm head segmentation, Neural Process. Lett. 54 (1) (2022) 537–557.

[33] S. Zou, C. Li, H. Sun, P. Xu, J. Zhang, P. Ma, Y. Yao, X. Huang, M. Grzegorzek, TOD-CNN: An effective convolutional neural network for tiny object detection in sperm videos, Comput. Biol. Med. 146 (2022) 105543.

[34] G. Cupples, M.T. Gallagher, D.J. Smith, J.C. Kirkman-Brown, Heads and tails: requirements for informative and robust computational measures of sperm motility, in: International Symposium on Spermatology, Springer, 2021, pp. 135–150.

[35] C. Leung, Z. Lu, N. Esfandiari, R.F. Casper, Y. Sun, Detection and tracking of low contrast human sperm tail, in: IEEE International Conference on Automation Science and Engineering, IEEE, 2010, pp. 263–268.

[36] A. Bijar, M. Mikaeili, R. Khayati, et al., Fully automatic identification and discrimination of sperm's parts in microscopic images of stained human semen smear, J. Biomed. Sci. Eng. 5 (2012) (2012).

[37] C. Dai, Z. Zhang, J. Huang, X. Wang, C. Ru, H. Pu, S. Xie, J. Zhang, S. Moskovtsev, C. Librach, et al., Automated non-invasive measurement of single sperm's motility and morphology, IEEE Trans. Med. Imaging 37 (10) (2018) 2257–2265.

[38] Z. Zhang, Robotic Manipulation and Selection of Single Sperm for in Vitro Fertilization (Ph.D. thesis), University of Toronto (Canada), 2019.

[39] A. Fraczek, G. Karwowska, M. Miler, J. Lis, A. Jezierska, M. Mazur-Milecka, Sperm segmentation and abnormalities detection during the ICSI procedure using machine learning algorithms, in: International Conference on Human System Interaction, 2022, pp. 1–6.

[40] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error measurement to structural similarity, IEEE Trans. Image Process. 13 (1) (2004).

[41] A. Rényi, On measures of entropy and information, in: Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, Vol. 4, University of California Press, 1961, pp. 547–562.

[42] K.-S. Chuang, H.-L. Tzeng, S. Chen, J. Wu, T.-J. Chen, Fuzzy c-means clustering with spatial information for image segmentation, Comput. Med. Imaging Graph. 30 (1) (2006) 9–15.

[43] Y. Wang, J. Riordon, T. Kong, Y. Xu, B. Nguyen, J. Zhong, J.B. You, A. Lagunov, T.G. Hannam, K. Jarvi, et al., Prediction of DNA integrity from morphological parameters using a single-sperm DNA fragmentation index assay, Adv. Sci. 6 (15) (2019) 1900712.

[44] C. McCallum, J. Riordon, Y. Wang, T. Kong, J.B. You, S. Sanner, A. Lagunov, T.G. Hannam, K. Jarvi, D. Sinton, Deep learning-based selection of human sperm with high DNA integrity, Commun. Biol. 2 (1) (2019) 1–10.

[45] J.B. You, C. McCallum, Y. Wang, J. Riordon, R. Nosrati, D. Sinton, Machine learning for sperm selection, Nat. Rev. Urol. 18 (7) (2021) 387–403.

[46] Z. Zhang, J. Liu, J. Meriano, C. Ru, S. Xie, J. Luo, Y. Sun, An automated system for investigating sperm orientation in fluid flow, in: International Conference on Robotics and Automation, IEEE, 2016, pp. 3661–3666.

[47] G. Saggiorato, L. Alvarez, J.F. Jikeli, U.B. Kaupp, G. Gompper, J. Elgeti, Human sperm steer with second harmonics of the flagellar beat, Nat. Commun. 8 (1) (2017) 1–9.

[48] P. Hernandez-Herrera, F. Montoya, J.M. Rendón-Mancha, A. Darszon, G. Corkidi, 3-D+t human sperm flagellum tracing in low SNR fluorescence images, IEEE Trans. Med. Imaging 37 (10) (2018) 2236–2247.

[49] M.T. Gallagher, G. Cupples, E.H. Ooi, J. Kirkman-Brown, D. Smith, Rapid sperm capture: high-throughput flagellar waveform analysis, Hum. Reprod. 34 (7) (2019) 1173–1185.

[50] B.J. Walker, S. Phuyal, K. Ishimoto, C.-K. Tung, E.A. Gaffney, Computer-assisted beat-pattern analysis and the flagellar waveforms of bovine spermatozoa, R. Soc. Open Sci. 7 (6) (2020) 200769.

[51] J.N. Hansen, S. Rassmann, J.F. Jikeli, D. Wachten, SpermQ–A simple analysis software to comprehensively study flagellar beating and sperm steering, Cells 8 (1) (2018) 10.

[52] R. Polikar, Ensemble based systems in decision making, IEEE Circuits Syst. Mag. 6 (3) (2006) 21–45.

[53] J. Luengo, R. Moreno, I. Sevillano, D. Charte, A. Peláez-Vegas, M. Fernández-Moreno, P. Mesejo, F. Herrera, A tutorial on the segmentation of metallographic images: Taxonomy, new metaldam dataset, deep learning-based ensemble model, experimental analysis and challenges, Inf. Fusion 78 (2022) 232–253.

[54] T. Dang, T.T. Nguyen, J. McCall, E. Elyan, C.F. Moreno-García, Two layer ensemble of deep learning models for medical image segmentation, 2021, arXiv preprint arXiv:2104.04809.

[55] D.H. Wolpert, Stacked generalization, Neural Netw. 5 (2) (1992) 241–259.

[56] I. Sirazitdinov, M. Kholiavchenko, T. Mustafaev, Y. Yixuan, R. Kuleev, B. Ibragimov, Deep neural network ensemble for pneumonia localization from a large-scale chest x-ray database, Comput. Electr. Eng. 78 (2019) 388–399.

[57] T. Gabruseva, D. Poplavskiy, A. Kalinin, Deep learning for automatic pneumonia detection, in: IEEE Conference on computer vision and pattern recognition workshops, 2020, pp. 350–351.

[58] M.M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A.R. Rudnicka, C.G. Owen, S.A. Barman, An ensemble classification-based approach applied to retinal blood vessel segmentation, IEEE Trans. Biomed. Eng. 59 (9) (2012) 2538–2548.

[59] S. Seferbekov, V. Iglovikov, A. Buslaev, A. Shvets, Feature pyramid network for multi-class land segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 272–275.

[60] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: Convolutional block attention module, in: European Conference on Computer Vision, 2018, pp. 3–19.

[61] A.J. Larrazabal, C. Martínez, J. Dolz, E. Ferrante, Orthogonal ensemble networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 594–603.

[62] B.C. Wallace, I.J. Dahabreh, Improving class probability estimates for imbalanced data, Knowl. Inf. Syst. 41 (1) (2014) 33–52.

[63] J.N. Hansen, A. Gong, D. Wachten, R. Pascal, A. Turpin, J.F. Jikeli, U.B. Kaupp, L. Alvarez, Multifocal imaging for precise, label-free tracking of fast biological processes in 3D, Nature Commun. 12 (1) (2021) 4574.

[64] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[65] Y. Sun, K.P. Pranav Shenoy, J. Shimamura, A. Sagata, Concatenated feature pyramid network for instance segmentation, in: IEEE Fifth International Conference on Multimedia Big Data, IEEE, 2019, pp. 297–301.

[66] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, et al., Hybrid task cascade for instance segmentation, in: IEEE Conference on computer vision and pattern recognition, 2019, pp. 4974–4983.

[67] O. Ronneberger, P. Fischer, 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.

[68] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, Y. Bengio, The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 11–19.

[69] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[70] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, UNet++: A nested U-Net architecture for medical image segmentation, in: Deep learning in medical image analysis and multimodal learning for clinical decision support, Springer, 2018, pp. 3–11.

[71] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, et al., Attention U-Net: Learning where to look for the pancreas, 2018, arXiv preprint arXiv:1804.03999.

[72] T. Fan, G. Wang, Y. Li, H. Wang, MA-Net: A multi-scale attention network for liver and tumor segmentation, IEEE Access 8 (2020) 179656–179665.

[73] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, C. Fan, SA-Unet: Spatial attention U-net for retinal vessel segmentation, in: International Conference on Pattern Recognition, IEEE, 2021, pp. 1236–1242.

[74] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: The European conference on computer vision, 2018, pp. 801–818.

[75] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, J. Wu, UNet 3+: A full-scale connected UNet for medical image segmentation, in: IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2020, pp. 1055–1059.

[76] H. Wang, P. Cao, J. Wang, O.R. Zaiane, UcTransNet: rethinking the skip connections in U-Net from a channel-wise perspective with transformer, in: AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 2441–2449.

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

[78] E.S. Uysal, M.Ş. Bilici, B.S. Zaza, M.Y. Özgenç, O. Boyar, Exploring the limits of data augmentation for retinal vessel segmentation, 2021, arXiv preprint arXiv:2105.09365.

[79] K. Zhao, Q. Han, C.-B. Zhang, J. Xu, M.-M. Cheng, Deep hough transform for semantic line detection, IEEE Trans. Pattern Anal. Mach. Intell. 44 (9) (2021) 4793–4806.

[80] J. Yi, S. Yoon, Patch SVDD: Patch-level SVDD for anomaly detection and segmentation, in: Proceedings of the Asian Conference on Computer Vision, 2020.

[81] J. Qi, Y. Gao, Y. Hu, X. Wang, X. Liu, X. Bai, S. Belongie, A. Yuille, P. Torr, S. Bai, Occluded video instance segmentation: A benchmark, Int. J. Comput. Vis. (2022).

[82] P. Voigtlaender, J. Luiten, P.H. Torr, B. Leibe, Siam R-CNN: Visual tracking by re-detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 6578–6588.