



Real-like synthetic sperm video generation from learned behaviors

Sergio Hernández-García¹ · Alfredo Cuesta-Infante¹ · Dimitrios Makris² · Antonio S. Montemayor¹

Accepted: 23 February 2025
© The Author(s) 2025

Abstract

Computer-assisted sperm analysis is an open research problem, and a main challenge is how to test its performance. Deep learning techniques have boosted computer vision tasks to human-level accuracy, when sufficiently large labeled datasets were provided. However, when it comes to sperm (either human or not) there is lack of sufficient large datasets for training and testing deep learning systems. In this paper we propose a solution that provides access to countless fully annotated and realistic synthetic video sequences of sperm. Specifically, we introduce a parametric model of a spermatozoon, which is animated along a video sequence using a denoising diffusion probabilistic model. The resulting videos are then rendered with a photo-realistic appearance via a style transfer procedure using a CycleGAN. We validate our synthetic dataset by training a deep object detection model on it, achieving state-of-the-art performance once validated on real data. Additionally, an evaluation of the generated sequences revealed that the behavior of the synthetically generated spermatozoa closely resembles that of real ones.

Keywords Sperm modeling · Sperm analysis · Diffusion models · Synthetic video · Synthetic dataset · Style transfer

1 Introduction

Semen evaluation is a fundamental diagnostic tool for assessing male fertility. In humans, it is utilized in a range of clinical contexts, including assisted reproduction, post-vasectomy monitoring, and the detection of sexually transmitted infections (STIs) [1–3]. In the livestock industry, it is widely used as tool for selection and breeding purposes [4–6].

Semen evaluation involves the assessment of various parameters by macroscopic and microscopic analysis of the samples. Macroscopic analysis measures characteristics such as semen composition, volume, liquefaction, viscosity, and pH. Microscopic evaluation involves the study of sperm clumping, motility, vitality, and cell count [7].

Automated systems to perform such microscopic tests have been developed and commercialized due to their high demand for three reasons. Firstly, qualified experts are needed, but they are scarce. Secondly, there are several dozens of spermatozoa moving in the semen sample, making it a difficult and time-consuming task. This entails a low upper limit to the expert's bandwidth. And thirdly, the subjectivity and lack of a traceable standard have been identified as causes of the high level of uncertainty associated with manual sperm motility and morphology assessments [3, 8]. However, laboratories are reluctant to accept the results until they have been validated against a reliable standard. In other words, the automated report must be compared to the human report, which is prone to a high level of uncertainty unless a great deal of expert effort and time is invested [9].

The arrival of deep learning has enabled computer vision systems to match or surpass human performance in multiple object detection and tracking, classification, recognition, segmentation and image generation in many fields [10–16]. Computer vision techniques have been applied in the specific context of microscopic semen analysis to morphological classification of sperm heads [17, 18], and to motility analysis. This process entails the detection and tracking of spermatozoa heads within highly concentrated samples [12, 19, 20]. Recent studies indicate that additional

✉ Sergio Hernández-García
sergio.hernandez@urjc.es

Alfredo Cuesta-Infante
alfredo.cuesta@urjc.es

Dimitrios Makris
d.makris@kingston.ac.uk

Antonio S. Montemayor
antonio.sanz@urjc.es

¹ Universidad Rey Juan Carlos, Móstoles, Spain

² Kingston University, London, UK

parameters, such as acrosome morphology (a cap-like structure on the sperm head) [21, 22] and beating patterns, may be important indicators of sperm health [23–28]. However, the aforementioned analyses necessitate images with a low concentration of spermatozoa. Furthermore, in the case of flagellum analysis, current state-of-the-art techniques also require the selection of the spermatozoon from the seminal fluid or its isolation in a video [23, 24, 29].

This paper presents a methodology that employs computer vision and deep learning techniques to provide unlimited, synthetic, video-realistic, fully labeled and on-demand video datasets that can be used for training, improving, and benchmarking computer-assisted sperm analysis (CASA) systems.

We propose a spline-based parametric model of a spermatozoon and use a Denoising Diffusion Probabilistic Model (DDPM) to animate it. To achieve this, we extract clean video sequences of isolated spermatozoa and model them according to the spline-based model. We then use the DDPM to learn the behavior of various classes of spermatozoa, which subsequently serves to generate new synthetic trajectories. Finally, we apply a Cyclic Generative Adversarial Network (GAN) to perform domain transfer on the animated trajectories, thereby obtaining realistic synthetic video sequences. In addition to this method, we provide a new fully annotated synthetic dataset for motility analysis. This dataset comprises tracking and detection labels, spline parameters modeling the flagellum shape, and motility category of each sperm cell.

2 Related works

In this section, we provide an overview of relevant topics to our case study, including recent proposals for CASA systems, available sperm datasets, techniques for generating synthetic sperm datasets, and applications of diffusion models in trajectory generation.

2.1 Computer-assisted sperm analysis

The computer-assisted sperm analysis (CASA) has been a research topic for a long time. According to [11], sperm analysis tools up to 2017 still delivered poor quality trajectories for high sperm concentration samples. Thus, during the last decade, new CASA proposals leveraged the emergence of deep learning for classification, detection and segmentation. Here, we mention some of the most recent.

A CNN was used for morphology classification of sperms in [30]. Similarly, a healthy sperm classification given only sperm head images was presented in [18]. In the same line, [17] carried out morphology classification by training a CNN in three openly available sperm morphology datasets. Segmentation of the sperm heads on the SCIAN-SpermGS dataset [31] was addressed in [15]. A sperm tracking method

suitable for samples recorded by a smartphone as a portable and low-cost platform was proposed in [32]. MotilitiAI consisted of a sperm analysis pipeline, including a tracker aimed at obtaining statistics of sperm movements, and a classification model to predict the fertility of the sample [33]. DeepSperm was a real time bull sperm detection trained with a dataset manually annotated by two experts [19]. Finally, in [12] YOLO was utilized to detect spermatozoa.

2.2 Sperm datasets

Although there are many open datasets of sperm video sequences, there is no homogeneity in terms of dataset size, capture conditions, microscope magnifications, colors and tones, annotations, and even classes in the ground truth. For example, the HuSHeM dataset [34] consists of 216 sperm images with sperm heads labeled as *normal*, *tapered*, *pyriform* and *amorphous*; whereas SMIDS dataset [30] collects up to 3000 images; however, these are only labeled as *normal* and *abnormal*. Below, we cite recent datasets to expose the diversity and lack of standardization.

- **VISEM** dataset [35] contains videos from 85 human patients of around 2 minutes, each recorded with a 400× magnification. These videos are annotated with sperm motility, sperm concentration, total sperm count, ejaculated volume, sperm morphology and sperm vitality among data related to fatty acids, fatty serum, sex hormones and anonymous participant related info as age. Thus, they are meant for classification or regression rather than object detection or tracking. To overcome this issue, VISEM-Tracking was released [36] with 20 videos of 30 seconds each.
- **SVIA** sperm detection and tracking dataset [37] has 101 human sperm video sequences obtained with WLJY-9000 CASA with ×20 magnification and ×20 electronic glasses. It was released split in three subsets: 1) for object detection tasks, with 3590 images and 125K annotated objects in three classes (sperm, debris or leukocyte), 2) for sperm head segmentation and tracking, with 451 frames and 26K annotations, and 3) head morphology classification with 125K images
- **SeSVID** dataset [38] contains 12 videos of human semen recorded under 100× magnification lens. The purpose of this dataset is solving object detection tasks providing 92,329 labeled objects (68,244 are sperm) in 1175 images. Unlike other datasets in which the ground truth bounding boxes frame only the sperm's heads, here they cover the entire head and tail.
- **SCIANSpermSegGS** dataset [15] is specifically meant for segmenting the different parts inside the head (head, acrosome, cell nuclei and midpiece).

Recent works highlights the importance of the movement of the flagellum and suggests that it should be taken into account in fertility and motility assessments [23–28].

2.3 Synthetic datasets

Two different approaches to synthetic data generation in the domain of sperm analysis have been proposed in the literature. The first one involves modeling the sperm using expert knowledge in the field while the second is based on the use of generative AI techniques.

Modeling the sperm morphology and motility aims to overcome such a heterogeneity and obtain fully annotated datasets at the same time.

In the context of motility analysis, two approaches [39, 40] have been applied to generating video sequences of schematic spermatozoa, modeling the spatial movement of the head and flagellum. These approaches analyzed different kinds of spermatozoon movements [7], and developed a mathematical model of those. However, these models only partially capture the nonlinear complexity of sperm movement, which can be influenced by various movement regimes, flagellum elasticity, and medium viscosity [27]. Regarding the realism of the generated videos, in the best cases, predefined hand-made rules over schematic frames are employed, such as adding Gaussian blur and salt-and-pepper noise [39]. Additionally, other authors incorporated a background extracted from real videos and included floating particles, such as white circles and shadows, to enhance realism [41].

The second line of research focuses on data driven techniques for generating synthetic images. Specifically, these works rely in Generative Adversarial Networks (GAN) [42] to augment a given dataset. In this context, GANs have been trained in [43–45] to augment the number of images from SMIDS, HuSheM, SCIAN-MorphoSpermGS, MHSMA as well as one private dataset [44]. In [13] a GAN with capsule network architecture was used to augment and balance the HuSheM dataset. However, all these GAN-based approaches are focused on morphological analysis of the head only, but none of them provide segmentation of the whole sperm (head + flagellum).

2.4 Denoising diffusion probabilistic models

Denoising Diffusion Probabilistic Models [46], *Diffusion Models* for short, have shown to be more robust than earlier generative architectures, including Variational Autoencoders (VAE) [47], GAN [42] and Deep Autoregressive Models [48].

Their applications span a diverse range of tasks, such as conditional image generation [49], image to image translation [50], text to image generation [16], point clouds

completion [51], natural language processing [52], and time series forecasting [53] among others [54].

The use of diffusion models for trajectory generation (related to the scope of this paper) was proposed in [55]. In this approach, trajectory generation is formulated as an inpainting process where the diffusion model transforms a random trajectory into a feasible and consistent one between two known locations. Similar approaches have been later successfully applied to modeling vehicle trajectories [56] and animating humanoid avatars [57–59] but none of them have been use on spermatozoa data.

3 Sperm video generation

Sperm video generation refers to the process of delivering a fully annotated sequence of frames that mimics both the dynamics of the sperm flagellum over time and the look and feel of the original video in terms of lighting, noise, artifacts, etc.

To this end, we firstly identify those *usable* spermatozoa from a video sequence and then carry out preprocessing on each one independently that results in a vectorized representation. The information extracted is then used to fit a parametric representation for each spermatozoon in each frame. Next, a model to generate new representations is learned from those collected from the real videos with a Denoising-Diffusion Probabilistic model. Finally, the representations generated are embedded into video frames and a style transfer neural network renders the resulting *schematic* sperm in pitch black background into a realistic video sequence. The details of each step are given in the following sections.

3.1 Spermatozoon windowing

By windowing, we refer to the process of extracting every single spermatozoon from a video sequence that complies with certain requirements, and processing it to deliver a cropped, segmented and rotated version of it and its neighborhood in a sequence of windows.

To begin with, a YOLO v5 [10] detection network is trained on a previously hand-annotated dataset to locate the head of each sperm cell in every frame. Besides, in order to preserve the identities of the detections, all of them are tracked throughout the video sequence using Kalman filtering. Thus, the result is a collection of N detections that we denote as $\{\delta^{(i)}\}$, for $i = 1 \dots N$; and such that

$$\delta^{(i)} = \left[(x_1^{(i)}, y_1^{(i)}), (x_2^{(i)}, y_2^{(i)}), \dots, (x_T^{(i)}, y_T^{(i)}), \text{remove}^{(i)} \right],$$

where $(x_t^{(i)}, y_t^{(i)})$, with $t = 1, \dots, T$, is the center of the bounding box with respect to the t -th frame and $remove^{(i)}$ is a boolean flag for the i -th detection initialized to False. In other words, we use the term “ i -th spermatozoon” to refer to the collection $\delta^{(i)}$ that univocally represents the position of the spermatozoon with identity i along the video sequence of T frames, along with the boolean flag.

In video of sperm, scenes are cluttered with frequent crossovers, clustering and splitting. For the method proposed in this paper, we need spermatozoa that do not encounter others during the whole sequence in its neighborhood. Let $\mathcal{B}_t((x_t^{(i)}, y_t^{(i)}); r)$, a ball of radius r centered in $(x_t^{(i)}, y_t^{(i)})$, be the neighborhood of the i -th spermatozoon in the t -th frame of the video. If there exists a j -th spermatozoon at the same frame t such that $(x_t^{(j)}, y_t^{(j)}) \in \mathcal{B}_t$, then $remove^{(i)} = True$. Hence, after checking all the possible crossings $(\delta^{(i)}, \delta^{(j)})$, we filter out those with the flag $remove = True$. Hence, the remaining are *usable*, as referred to above.

Let $\{\zeta^{(i)}\} = [(x_1^{(i)}, y_1^{(i)}), (x_2^{(i)}, y_2^{(i)}), \dots, (x_T^{(i)}, y_T^{(i)})]$, for $i = 1, \dots, n$, be the usable spermatozoa detections reindexed from 1 up to $n \leq N$ and without the flag $remove$ because it is no longer needed. Such a reindexing does not affect to the terminology introduced, so in the following “ i -th spermatozoon” is $\zeta^{(i)}$.

Thus, for each $\{\zeta^{(i)}\}$ and each t , the following steps are performed:

1. Crop a squared window centered at $(x_t^{(i)}, y_t^{(i)})$ of size $S \times S$.

Hence, the center of the window is locally (i.e. within the window) referenced as $(0, 0)$ but globally (i.e. within the video frame) as $(x_t^{(i)}, y_t^{(i)})$.

2. Segment the spermatozoon in the window. To this end, adaptive threshold, dilate and erode operations are used.
3. Fit an ellipse to the full spermatozoa using the Fitzgibbon algorithm [60].
4. Rotate the window so that the major axis of the ellipse becomes horizontal and the spermatozoon is heading to the right. Let α be the angle of that rotation, then we define a unit vector $\vec{\alpha} = (\alpha_x, \alpha_y)$ such that

$$\alpha_x = 1/\sqrt{1 + \tan^2 \alpha} \quad \text{and} \quad \alpha_y = \tan \alpha / \sqrt{1 + \tan^2 \alpha}. \tag{1}$$

We adopt this vector to facilitate further learning and inference processes.

These steps are depicted in Fig. 1(a)-(c), together with other information described in the next subsection. Notice that the center of the window is the center of the detection, but not necessarily the center of the head.

3.2 Spermatozoon parametric representation

We propose a parametric representation of the i -th spermatozoon in frame t in terms of its head, flagellum, full body

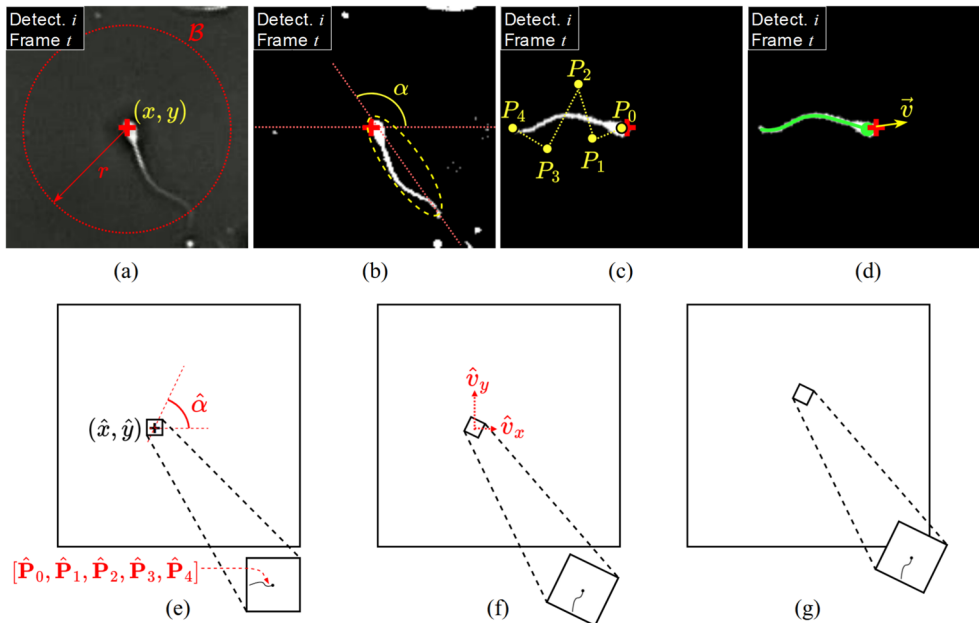


Fig. 1 Top row: Process to obtain the parametric representation of a spermatozoon; (a) Neighborhood \mathcal{B} of a spermatozoon, (b) Rotation, (c) Spline fitting, (d) velocity vector. Bottom row: Trajectory generation

of a schematic spermatozoon; (e) insert the window into the frame and rotate it back, (f) add the velocity vector to the position. (g) position and rotation in the next frame

and velocity. For the sake of clarity, we omit superscript (*i*) and subscript *t* in all the parameters, unless it is needed.

- **Head:** For the sake of simplicity, the head is just a circle of the same size in all the windows, locally centered in (0, 0) and globally in (*x*, *y*).
- **Flagellum:** We propose to model the flagellum as a 4-order Bezier Spline, given by:

$$B(\lambda; \mathcal{P}^{(4)}) = \sum_{k=0}^4 \binom{4}{k} (1-\lambda)^{4-k} \lambda^k \mathbf{P}_k, \tag{2}$$

where $\mathcal{P}^{(N)} = \{\mathbf{P}_k\}$, for $k = 1, \dots, N$ and $\mathbf{P}_k \in \mathbb{R}^2$, are the control points; and λ generates the curve as it goes from 0 to 1. \mathbf{P}_0 and \mathbf{P}_4 are set to the head and end of the tail respectively, while $\mathbf{P}_1, \mathbf{P}_2$ and \mathbf{P}_3 are fitted by means of the Nelder-Mead algorithm. Control points are depicted in Fig. 1(c).

- **Body:** We use the unit vector $\vec{\alpha} = [\alpha_x, \alpha_y]$ as obtained in (1).
- **Velocity:** We use a velocity vector $\vec{v} = [v_x, v_y]$, where $v_x = (x_{t+1} - x_t)$ and $v_y = (y_{t+1} - y_t)$ to indicate the direction and magnitude of the *i*-th spermatozoon global translation in two successive frames. Vector \vec{v} is shown in Fig. 1(d).

All the representation parameters are summarized in Table 1, indicating whether they are obtained locally (with respect to the window) or globally (with respect to the frame).

Finally, we introduce the representation vector at frame *t* of the *i*-th spermatozoon

$$\rho_t = [\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4, v_x, v_y, \alpha_x, \alpha_y],$$

so the ordered array of ρ_t from $t = 1$ to $t = T$ is the fully parametric representation of the trajectory followed by *i*-th spermatozoon along a video sequence of *T* frames

$$\tau = [\rho_1, \rho_2, \dots, \rho_T].$$

Table 1 Summary of parameters of the spermatozoon representation

<i>i</i> -th spermatozoon at frame <i>t</i> -th	Parameter	Meaning (from global frame or local window)
Head	<i>x</i> , <i>y</i>	Detection centre (global)
Flagellum	$\mathbf{P}_0 \dots \mathbf{P}_4$	Control points of a 4-order Bezier spline (local)
Full body	α_x, α_y	Orientation of the spermatozoon (global)
Velocity	v_x, v_y	translation of the head from frame <i>t</i> to <i>t</i> + 1 (global)

Notice that trajectory vector τ renders both the movement and the travel of a schematic spermatozoon within the frame.

3.3 Sperm trajectory synthesis with diffusion models

Denoising-Diffusion Probabilistic Models (DDPM) [46, 55] have gained popularity due to the excellent performance in text-to-image generation. In a nutshell, given a sample (not necessarily an image) s_0 , it is successively corrupted by adding a small amount of Gaussian noise during *H* steps. In any two consecutive samples s_k and s_{k+1} , it is possible to train a neural network for denoising s_{k+1} in a supervised way, using s_k as ground truth. During the training process the indices of the samples are also input. In inference, after all the denoising steps, a clean sample is obtained from a fully noisy one. Additionally, in text-to-image tasks, the resulting image is conditioned by the caption.

In the context of this paper, the sample is the trajectory vector τ of the *i*-th spermatozoon along a video sequence of *T* frames. The goal is to produce a new trajectory $\hat{\tau}$ conditioned on the representation vector at the first frame of the sequence ρ_1 from a fully noisy vector of the same size and after *H* denoising steps by the neural network learned.

The process is depicted in Fig. 2; in which $\hat{\tau}_H$ consists of the true ρ_1 followed by a random array that completes the trajectory vector. After one step *h* this array is modified (denoised) towards more meaningful values. After *H* steps, the array has been transformed into *T* - 1 meaningful representation vectors, resulting into a generated trajectory $\hat{\tau}$. Then with $\hat{\tau}$ and an initial global position (*x*, *y*), a schematic spermatozoon is inserted in each frame of the video sequence. Finally, this process is repeated to produce *N* synthetic and fully labeled but schematic spermatozoa moving around.

3.4 Synthesis of real-looking videos

The resulting video sequence lacks of background noise, motion blur, lens and chromatic aberrations, artifacts, shadows,

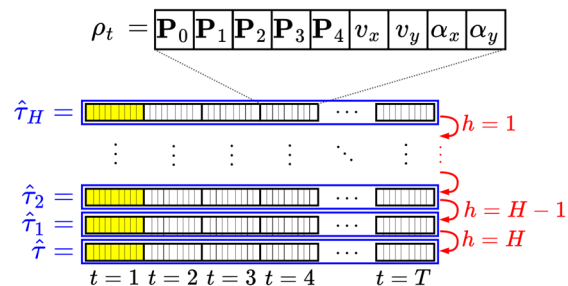


Fig. 2 Diffusion denoising process to generate a trajectory vector $\hat{\tau}$ in *H* steps. The first representation vector (with background filled) is always imposed to be the true ρ_1

and other particles. Besides, real spermatozoa do not have a circular head and the mid-piece is missing in the synthetic one.

To deliver a realistic video it is necessary to incorporate all these features. To this end, we propose to do style transfer with a CycleGAN [61]. The CycleGAN learns to transform a schematic frame (domain A) to the style of a real frame (domain B) and vice versa. Completing a cycle from domain A to B and back ensures structural consistency between the generated images. Besides, both domains are indeed unpaired datasets, so the training process requires no human supervision at all.

4 Experimental results

In this section we conduct exhaustive experiments to assess the morphology, motility and utility of the method proposed in quantitative terms. In addition, we present a qualitative analysis to validate the results from an expert perspective.

4.1 Dataset and experimental approach

We use a private porcine semen dataset provided by a commercial company that consists of 28 one-second videos recorded at 25 frames per second with resolution of 1280×1024 pixels [62]. The videos were taken with a Motic Panthera C2 microscope, with a magnification ratio of $10\times$, and a Blackfly USB3 camera with a 12.3MP Sony CMOS sensor.

After filtering out of all the spermatozoa initially detected, we end up with 72 usable sequences. Notice that the goal of this paper is to create synthetic video sequences using a suitable for the problem data augmentation method, instead of simplistic modifications such as sudden rotations or translations. We use a sliding interval of 16 frames, so a single sequence of 25 frames is transformed into 9 sequences of 16 frames (from frame 1 to 16, from 2 to 17, and so until from 9 to 24). The choice of 16 is a trade-off between the remaining sequence length and the number of times the dataset is increased. Frame 25 is necessary to have the velocity vector of the previous frame, but since there is not a following one, there is no sequence from frame 10 to 25. Hence, we end up with a total of 648 usable sequences with $T = 16$.

From each sequence we obtain its trajectory vector as defined in the previous section. Since the porcine spermatozoon are no longer than 70 pixels, the window size chosen is $S = 140$.

We categorize each one according to their trajectory, following the World Health Organization specifications [7] as “progressive” (46%) , “slowly progressive” (22%) and

“inmotile” (32%) sperm. We keep a stratified 10% of the sequences for testing and use the rest for training.

Our DDPM utilizes a U-Net [14] architecture with 3.96M of parameters. It incorporates a sinusoidal positional encoder [63] to determine the current step in the denoising process. The DDPM is applied to an input noisy trajectory of length 16, and uses 20 denoising steps. To further enhance training stability, we employ a model with an Exponential Moving Average [64]. The CycleGAN used for style transfer utilizes two U-Net generators, each comprising 11.4M parameters. The two discriminators models consist of convolutional networks, each with 2.8M parameters. Note that, once trained, only one U-Net generator is required for inference.

4.2 Baseline models

The following experiments consider three sets of synthetic sperm. A set derived from our method, and two sets derived from baseline models described below. We then compare their metrics with the real sperm set.

Gaussian model As a first baseline, we assume that the trajectory vectors are normally distributed; that is $\tau \sim \mathcal{N}(\mu, \Sigma)$, where \mathcal{N} stands for multivariate normal with mean μ and covariance matrix Σ computed on the training set. Thus, once the \mathcal{N} is fitted, new trajectory vectors $\hat{\tau}$ are just sampled from it.

LSTM model As a second baseline, we consider an LSTM model [65] in order to capture the time dependence within the sequence of frames. Specifically, a trajectory generated with this model would be $\hat{\tau} = [\hat{\rho}_1 = \rho_1, \hat{\rho}_2 = \rho_2, \hat{\rho}_3 = \rho_3, \hat{\rho}_4 = \rho_4, \hat{\rho}_5, \dots, \hat{\rho}_T]$, such that we compute $\hat{\rho}_t = \text{LSTM}(\hat{\rho}_{t-1}, \hat{\rho}_{t-2}, \hat{\rho}_{t-3}, \hat{\rho}_{t-4})$, for $t = 5, \dots, T$. In other words, to generate a trajectory requires a tuple of its first four true representation vectors.

4.3 Evaluation of synthetic sperm morphology

To assess the synthetic sperm morphology, we measure the similarity of the flagellum to real spermatozoa. To this end, we compare the respective distributions for each generated spline control point, $\Pr(\hat{\mathbf{P}}_0)$ to $\Pr(\hat{\mathbf{P}}_4)$, using the Kullback-Leibler (KL) divergence and the Earth mover’s distance (EMD). To compute the KL divergence we assume that all the distributions are univariate Gaussians. This is a strong assumption so we also compute the EMD approximating the continuous underlying distributions, $\Pr(\hat{\mathbf{P}}_j)$ and $\Pr(\mathbf{P}_j)$, by discrete PMFs. Since the shape of the flagellum is related to its motility, we consider the three categories namely progressive, slowly progressive and inmotile sperm. Additionally, to

estimate the distributions $\Pr(\hat{\mathbf{P}}_0)$ to $\Pr(\hat{\mathbf{P}}_4)$ for each type of sperm, we generate 1000 full trajectories per type.

The quantitative results are shown in Fig. 3, in which our method is always depicted as the dotted blue line with circle marks. In both top and bottom rows the lower, the better. Hence, any mark above the dotted line is worse than the method proposed here. Specifically, the top row shows that our method attains a KL divergence similar to the one between test and train set distributions (green diamonds). These results are confirmed in the bottom row, although the LSTM model fits better for slow sperm. However, LSTM requires four representation vectors in a row to generate the video sequence of slow spermatozoa while our method only requires one.

4.4 Evaluation of synthetic sperm motility

The evaluation of the sperm morphology takes into account the distribution of control points without considering the translation along the sequence. To assess the motility is equivalent to measuring the similarity of the trajectories generated with respect to those in the sequences of the test set. However, a generated sequence and a real sequence of progressive spermatozoa may be significantly different with respect to the path followed and still be similar in terms of in terms of distance traveled. In order to compare them, we generate 10 trajectories for each real trajectory of each type and use the Average Displacement Error (ADE) and the Min-

imum Average Displacement Error (MADE), which are two popular motion prediction metrics. Given that the generated trajectories are conditioned by the first frame, both generated and real trajectories begin at the same coordinates. Therefore, on average, the generated trajectories are not expected to diverge significantly from the real ones. However, ADE approaching zero is a warning of overfitting, since there is no generation but copycat of the true trajectory. Consequently, the only information needed is the coordinates $(x_t^{(i)}, y_t^{(i)})$ for every detected real spermatozoon $\zeta^{(i)}$ in every frame t , along with the generated ones $(\hat{x}_t^{(i)}, \hat{y}_t^{(i)})$.

The results are depicted in Fig. 4, which follows the same legend as Fig. 3; and again the lower, the better too. Thus, our proposed method (dotted blue line) performs better or equally well as the real video sequences and the baseline models on average considering ADE or MADE. The only exception is the evaluation of the slowly progressive sperm based on MADE. However, we stress that MADE takes into account only a single trajectory (the one with the minimum displacement error) per method, while the average across all trajectories (ADE) is similar for all methods for this sperm type. In addition, our method outperforms the baselines by much larger margin for the other two sperm types.

4.5 Qualitative evaluation

We present some qualitative results to assess the realism of the generated data in terms of morphology, motility and final rendering.

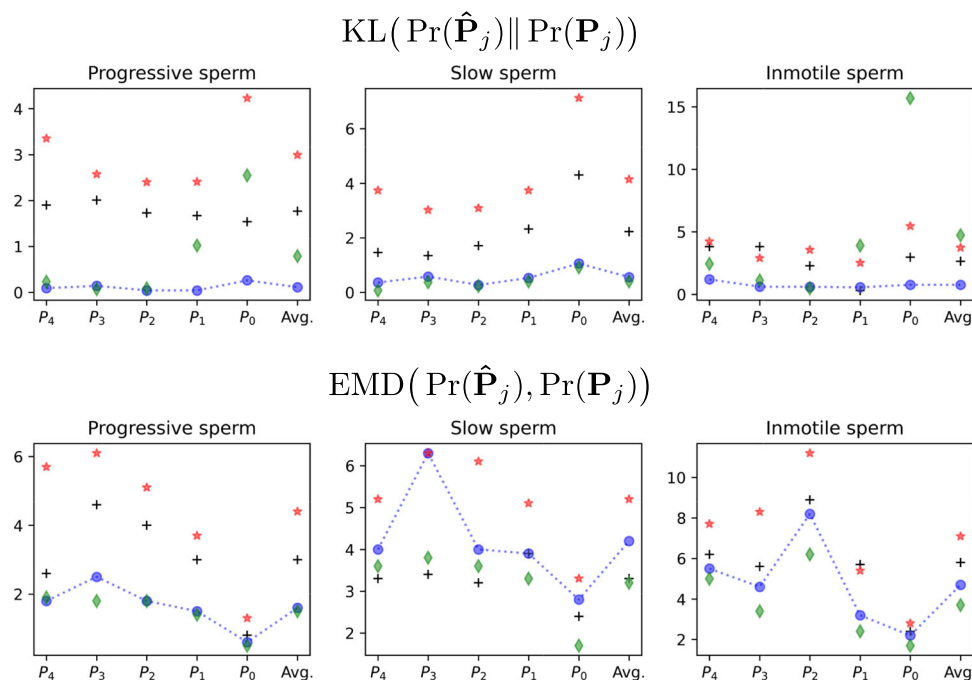


Fig. 3 Comparison of the distribution spline control points based on the KL divergence (Top) and the Earth Mover's Distance (Bottom). The lower, the better for both. The dotted line is the set derived from method proposed; +, * and ◇ represent LSTM, Gaussian and Train sets respectively

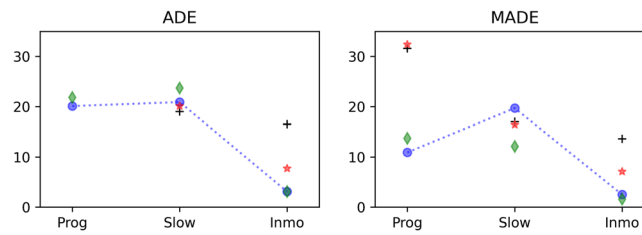


Fig. 4 (Left) Average Displacement Error, and (Right) Minimum Average Displacement Error; both with respect to the head of the spermatozoon. The lower, the better in both plots. The dotted line is the method proposed; +, ★ and ◇ represent LSTM, Gaussian and Train sets respectively

4.5.1 Morphology

With respect to the morphology, we show how the spline control points are distributed within the window in the upper row of Fig. 5, as well as the rendered spline in the lower row. For the sake of clarity and compactness, we only depict the progressive sperm generated by our method together with the training and test sets. Five clusters with a similar distribution are clearly visible in the three upper subplots, the only difference being the number of samples. Likewise, in the lower row, the distribution of rendered splines looks the same and it is consistent with the compared of progressive sperm.

Next, we show the morphology as the sperm travels through the frame in Fig. 6, in which two synthetic 8-frame sequences (middle and bottom rows) are confronted with a real sequence.

4.5.2 Motility

To illustrate the motility, we present 10 generated trajectories (solid blue) vs. 1 real trajectory (dotted red) for each type of sperm in Fig. 7. The upper row is the method proposed and the lower row is generated with the LSTM. According to the quantitative results in Fig. 4, both models perform

quite similar in terms of ADE. However, in Fig. 7, it can be appreciated that the proposed method produces longer and much more diverse paths than the LSTM.

4.5.3 Frames rendering

We compare the final rendering of the method with real frames in Fig. 8(a), in which the left plots are two samples of how a generated frame looks like before being transformed with the Cycle GAN, and the right plots are the outcomes of the Cycle GAN, hence synthetic frames. On the other hand, Fig. 8(b) shows how the cycle GAN transforms real frames (left) into schematic frames that seem to be extracted from the domain of the generated frames (right). Further qualitative comparisons of the other types of sperm are given in the Appendix.

4.5.4 Expert evaluation

A survey was conducted to assess the ability of experts to distinguish between real and synthetic images and videos. We recruited 19 participants: 3 from medicine and biology, 7 from the computer vision community, 7 with an interdisciplinary profile including both areas, and 2 from other

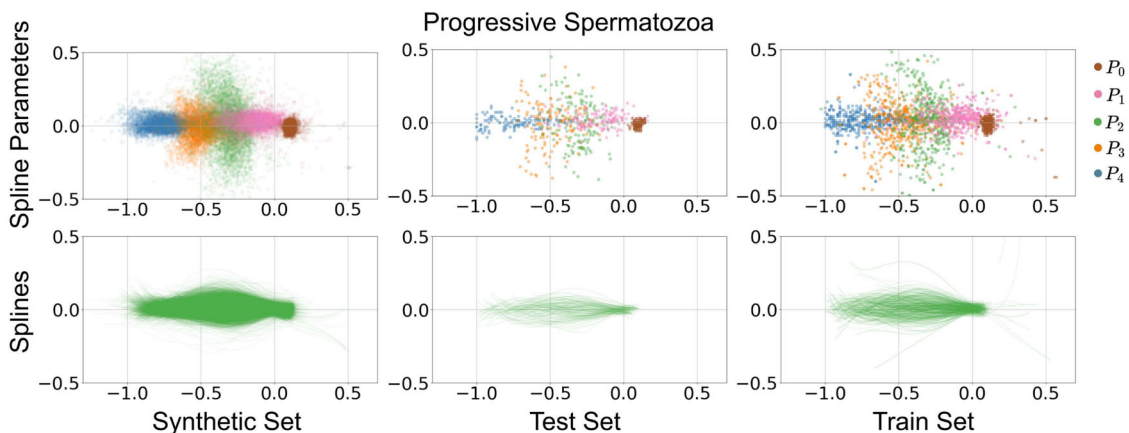


Fig. 5 Distributions of spline control points and respective flagellum for progressive spermatozoa in the generated synthetic set, test set and train set

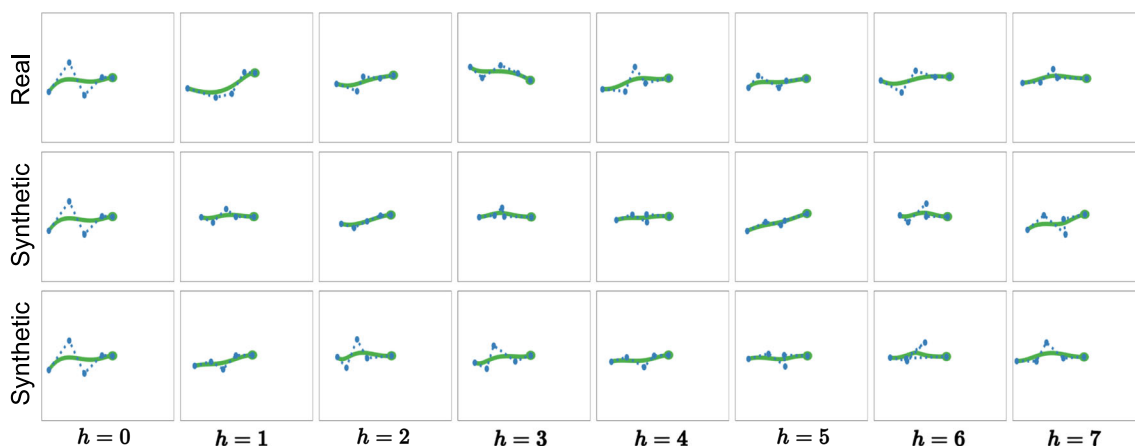


Fig. 6 A sample of single spermatozoon generation in the spline model domain. The upper row shows a 8-frame sequence extracted from the real video, the bottom two rows show two different generated sequences

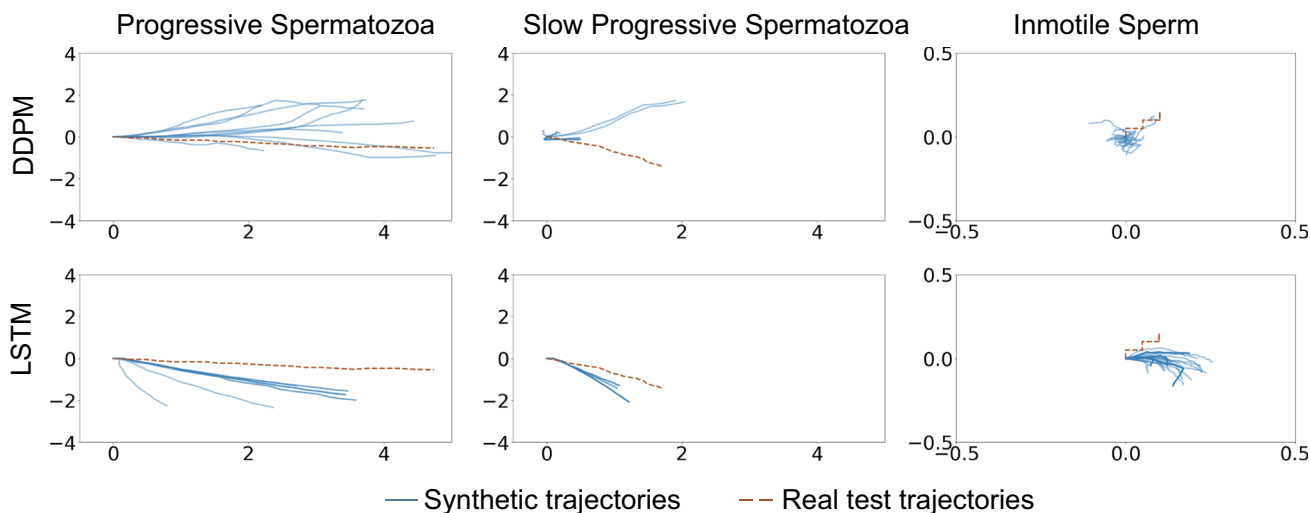
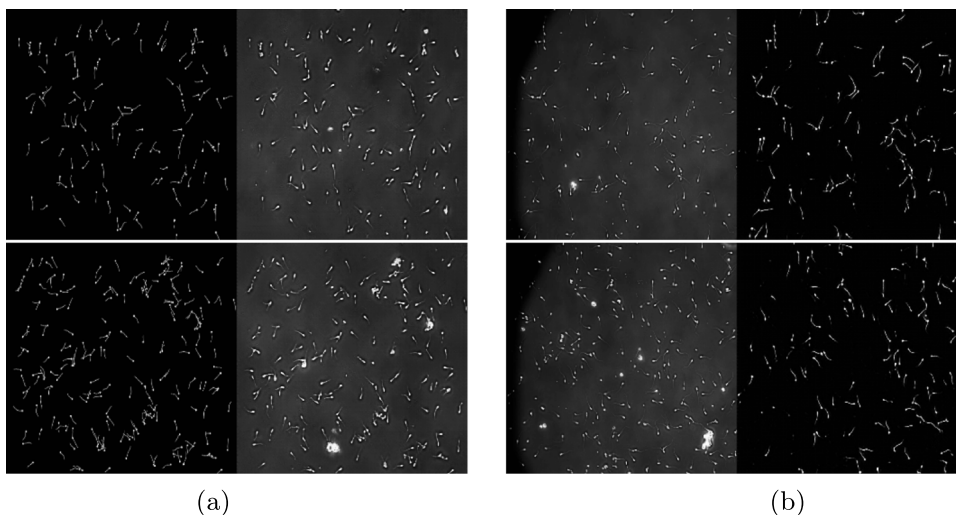


Fig. 7 Ten generated trajectories (solid blue) vs. one real trajectory (dotted red) for each type of sperm

Fig. 8 Cycle-consistency between schematic domain and true domain. (a) From the generated schematic to the generated real-looking frame. (b) From the true frame to its version in the schematic domain



fields. 16 out of 19 reported experience working with microscopic images, and 7 out of 19 with sperm images. Therefore, we categorized participants into two groups: sperm imaging specialists, referred to as *Qualified Experts*, and *Other Experts*. Each respondent was presented with 6 images, 3 real and 3 synthetic, one after the other and randomly ordered; and was asked to indicate to what extent they considered each image to be real or synthetic according to the following excluding options: *synthetic*, *likely synthetic*, *unsure*, *likely real*, or *real*.

To analyse the survey, we use the response to each image as a proxy of the likelihood that the image viewed belongs to a real video and assign the following values:

$$p(y = \text{'Real'}|x) = \begin{matrix} \text{synthetic} & \text{likely synthetic} & \text{unsure} & \text{likely real} & \text{real} \\ 0.01 & 0.25 & 0.5 & 0.75 & 0.99 \end{matrix}$$

Thus, each expert is treated as a classifier returning $p(y = \text{'Real'}|x)$ and a confusion matrix and ROC curve is obtained. The averaged results are shown in Fig. 9. The first remark is that both qualified experts and the rest of participants attain similar scores. The ROC curve, shown in Fig. 9(a) is close to the diagonal (dotted line), indicating that both groups perform similar to a random guess. The Precision, Recall and F1-Score metrics are shown in Fig. 9(b). With a confidence threshold of ≥ 0.75 , a precision value of 0.5 evidences that the half of the images classified as *real* and *likely real* by qualified experts were actually synthetic. As the threshold is raised to ≥ 0.99 (absolute certainty), qualified experts classified only 19% of real images correctly, while, the other experts identified just 6% of real images correctly. These results support the claim that synthetic images cannot be distinguished from real.

4.6 Application-related validation

Meaningful validation of the generated videos is challenging, as it requires assessing them within the context of the commercial CASA for which it is intended for. To overcome such a challenge, we propose training and using an object detector such as the YOLO v5 network to detect sperm on the two datasets:

1. a human-labeled dataset consisting of 42 frames that accounts for a total of 6938 spermatozoa.
2. a synthetic dataset consisting of 672 frames from 28 real-looking videos generated with the method proposed, which makes a total 117,331 spermatozoa.

We train three YOLO v5 models: 1) with 80% of the human-labeled dataset, 2) with 80% of the synthetic dataset, and 3) with the union of the two previous ones. Similarly, we test the above models with the remaining 20% of the human-labeled, synthetic and union of both datasets. The results are shown in Table 2.

4.6.1 Training YOLO-based sperm detection

Assuming that there is a scarce ‘human-labeled’ dataset and it is held for testing, can we use the synthetic dataset for training? The answer is obtained from the ‘Human-labeled’ column of Table 2. Training with synthetic data attains slightly higher F1-score than training with a human-labeled data (that we had kept for the sake of this comparison). The union of both training sets achieves a similar F1-score with the human-labeled test set. In other words, the synthetic

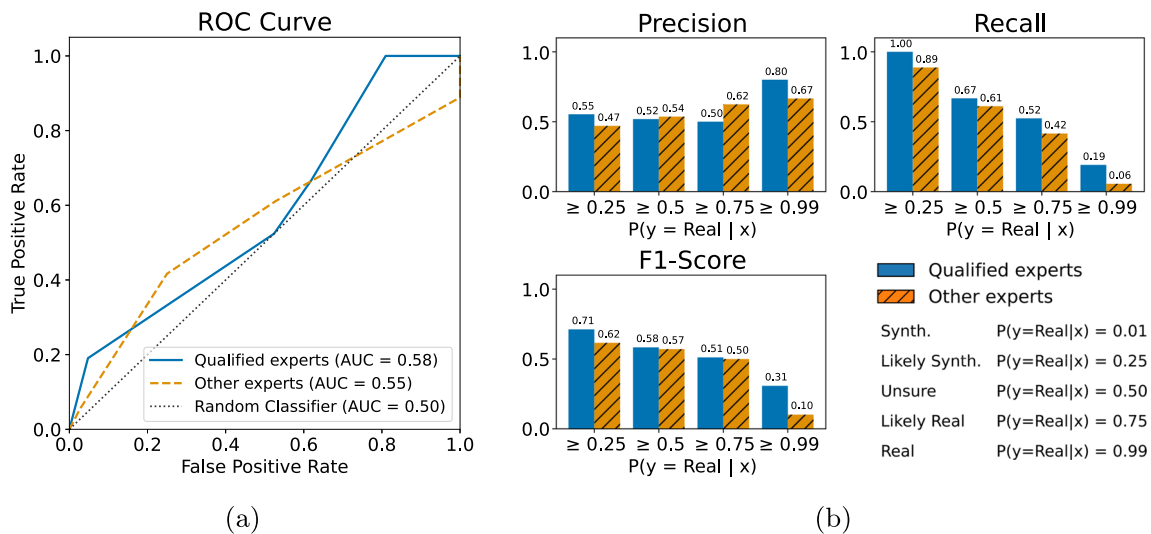


Fig. 9 Results of the classification of images as real by qualified and other experts: (a) ROC curve; (b) precision, recall and F1-score

Table 2 F1-score of three YOLO v5 models trained with human-labeled, fully annotated synthetic video (ours) and both

↓Train set \ Test set→	Human-labeled	Synthetic	Human + Synth.
Human-labeled	0.834	0.846	0.845
Synthetic	0.851	0.974	0.961
Human + Synth.	0.858	0.975	0.910

dataset by its own is capable of training a YOLO-based sperm detector at least as good as a human-labeled dataset, if the latter is not available.

4.6.2 Testing YOLO-based sperm detection

Assuming that YOLO-based sperm detector is a CASA whose performance we aim to evaluate, can we use the synthetic dataset for testing? We utilize synthetically generated videos, which constitute a fully annotated dataset, for evaluation purposes. Hence, we focus on the ‘Synthetic’ and ‘Human+Synth’ columns of Table 2. If the training set is human-labeled, the F1-score attained across all the three test sets is similar. However, the confidence in the results is significantly higher for ‘Synthetic’ and ‘Human+Synth.’ test sets, as both are 17 times larger than the ‘Human-labeled’ set. Conversely, if we train with a large fully annotated dataset (rows ‘Synthetic’ and ‘Human+Synth.’), the F1-score surpasses 90%.

5 Conclusions

We presented a novel framework for generating realistic synthetic videos of sperm. This framework aims to address limitations that currently hinder the integration of deep learning techniques into CASA systems. Our approach generates labeled videos, including head and flagellum morphology annotations, which can serve as a surrogate for human-labeled data. These synthetic videos have been shown to be useful for training and evaluating CASA systems relying on deep learning detection networks.

A parametric spermatozoon model was proposed to capture the key morphological and motility features of spermatozoa to simulate their trajectories across video frames. Subsequently, a Denoising Diffusion Probabilistic Model was utilized to learn spermatozoa behavior. This enabled the animation of the parametric model, thereby generating realistic motion patterns. These patterns were then embedded into frames to produce *schematic* video sequences. Finally, the generated videos were rendered with a realistic appearance using a CycleGAN for style transfer.

We conducted experiments to assess the goodness of the morphology, motility, appearance, and utility of the generated videos. The experimental evaluation confirmed that the proposed method generates spermatozoa trajectories that align with real data distribution.

Despite the encouraging results, our focus has remained on a private dataset utilized by a commercial CASA system. Extending its applicability to other datasets may require adaptations, such as a dataset-specific *spermatozoon windowing* procedure or manual annotation. This necessity arises due to the lack of openly available labeled datasets that meet the requirements of our method. These requirements include access to tracking labels for spermatozoon heads, and sperm concentrations that allow for the extraction of trajectories without intersections between spermatozoa. To the best of our knowledge, four currently available datasets provide sperm head detections [19, 36–38]. However, some of these datasets exhibit limitations, such as excessively high sperm cell concentrations [19, 37], low contrast, monochromatic images [38], or noisy and artifact-prone images [37]. Notably, none of the datasets provide annotations of the flagella. Despite these limitations, the core components of the method, namely the parametric spermatozoon model, the DDPM, and the CycleGAN, are not specifically related to a particular dataset or CASA. In fact, different videos only require adjustments in hyperparameters such as the window size, the neighborhood radius, the number of control points, the number of frames, etc.

The findings of this work indicate that synthetic data can serve as a valuable resource to support the evaluation of CASA systems and the training of deep learning algorithms for sperm analysis. Furthermore, these findings could potentially enable additional research in this field, such as conducting more detailed analyses of flagellar beating patterns, encouraging the development of new deep learning methodologies for sperm analysis, or serving as an educational resource for laboratory personnel.

Appendix A: Qualitative results

This section shows in Figs. 10 and 11 the distribution of the parameters and flagellum of the synthetic data compared to their respective training and test sets. The plots show the similarity among synthetic and real data.

This section includes more qualitative results regarding individual sperm sequences in Figs. 12 and 13. These figures show a comparison between a real sequence and two generated with a DDPM conditioned to an initial real state.

Finally, Fig. 14 shows the qualitative results of applying CycleGAN to perform a domain transfer from animated video sequences to real looking videos.

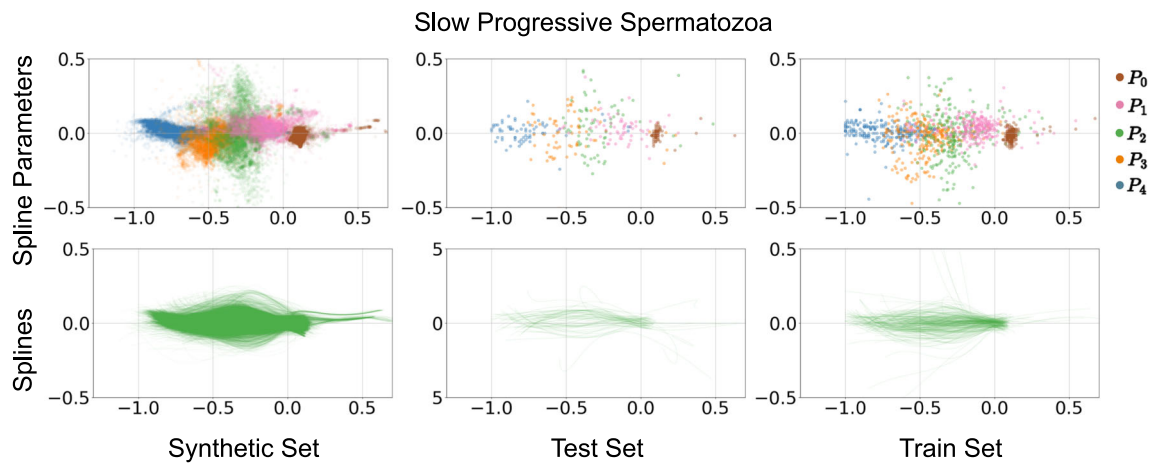


Fig. 10 Distributions of spline parameters and respective flagellum for slow progressive spermatozoa in the generated synthetic set, test set and train set

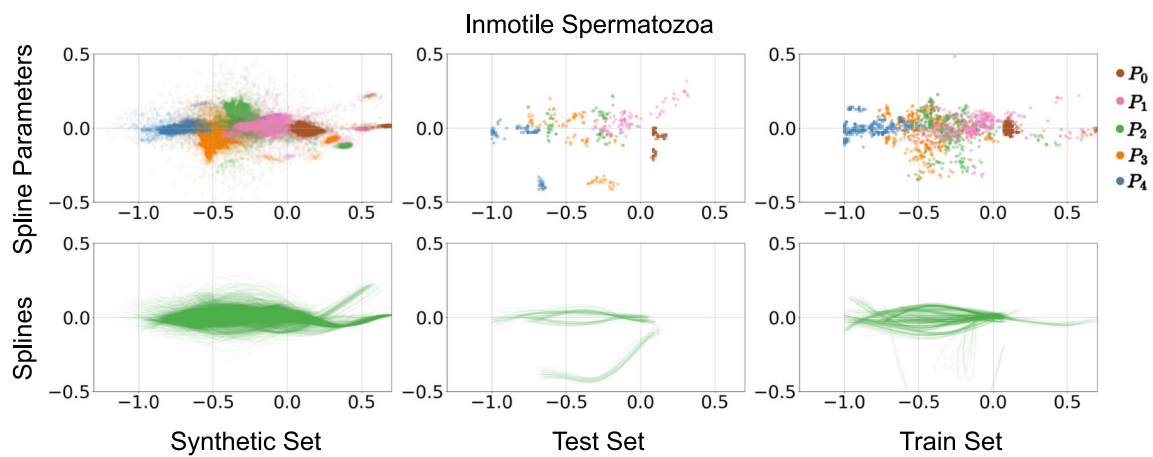


Fig. 11 Distributions of spline parameters and respective flagellum for immotile sperm in the generated synthetic set, test set and train set

Fig. 12 Samples of single spermatozoa generation within the parametric model. The upper row of each plot shows the reference sequence extracted from the real data, the bottom two rows show the sequences generated by the DDPM while it is conditioned to an initial and final state

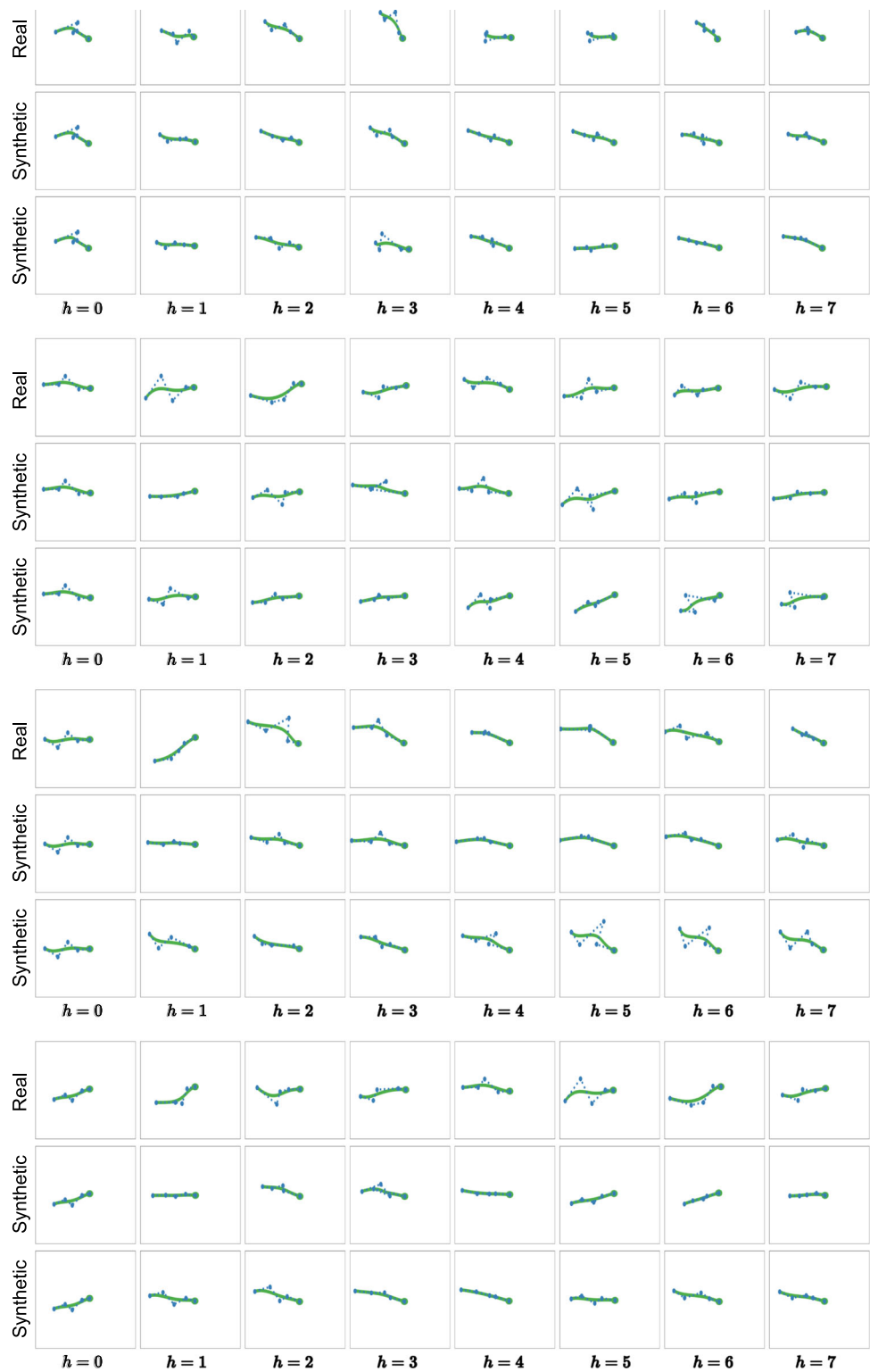


Fig. 13 A sample of single spermatozoon generation within the parametric model. Right column from each plot shows the reference sequence extracted from the real data, other columns show sequences generated by the DDPM while it is conditioned to an initial and final state

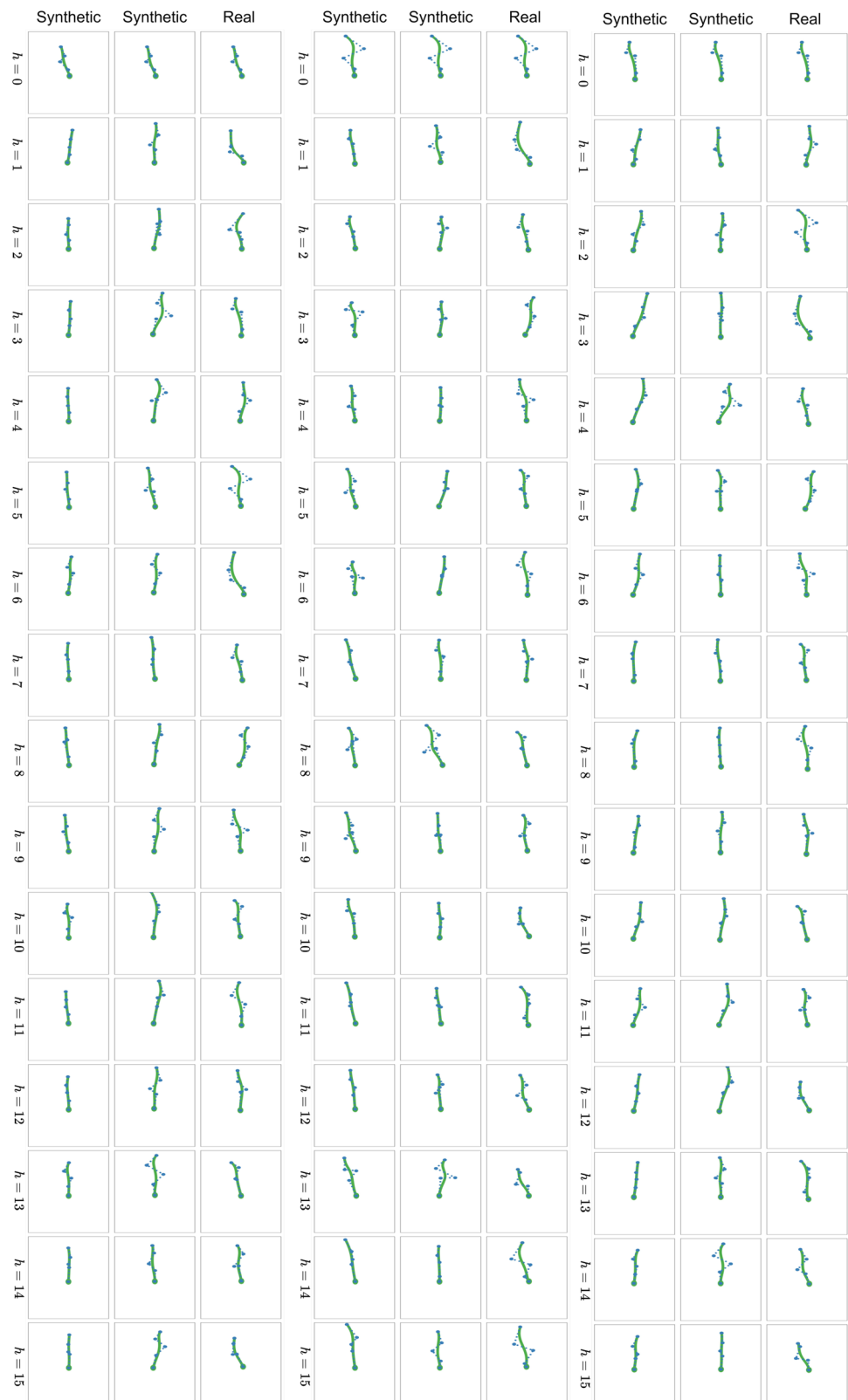
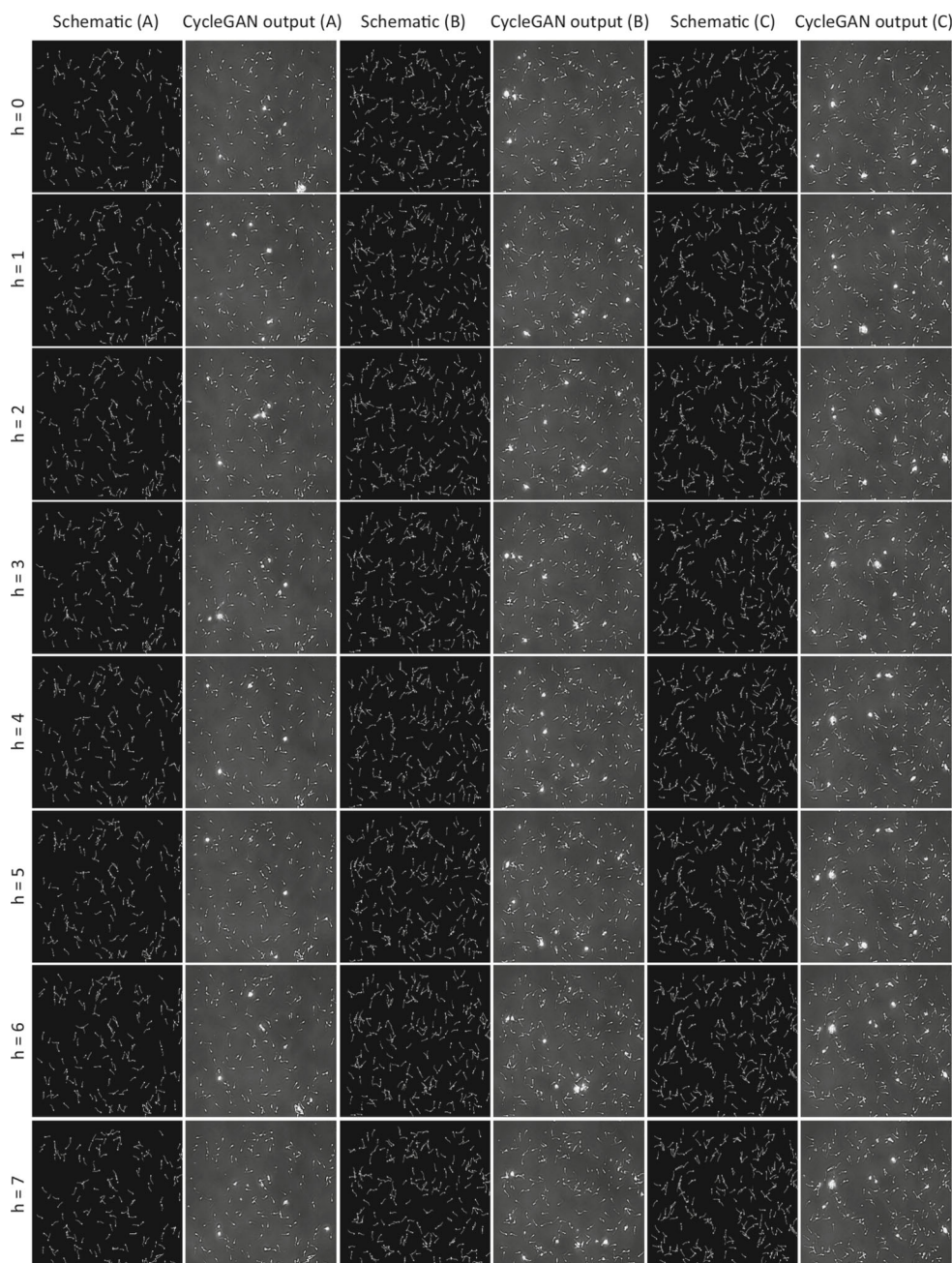


Fig. 14 Three synthetic video sequences with a horizon limited to 8 for visualization purposes. The schematic frames show animated spermatozoa. CycleGAN output frames show the respective frames after the domain transfer operation



Appendix B: Experimental details

This section includes specific details of the hyperparametrization of the models and experimental settings. Additional implementation details can be found in the associated GitHub repository: <https://github.com/SergioHdezG/sperm-diffuser>.

Parametric model normalization All methods used to generate parametric spermatozoa are trained on normalized trajectories. We separated global parameters from local parameters and normalized them in different ways. First, the global parameters refer to the spermatozoon’s movement

across a frame. These include body orientation ($\vec{\alpha}$) and velocity (\vec{v}). We rotate these vectors for every frame in each trajectory an angle given by

$$\theta_{t_0,i} = -\arctan\left(\frac{v_x}{v_y}\right)_{t_0,i},$$

where i represent the trajectory index, t_0 specifies that the first frame from trajectory i is used to calculate θ , and (v_x, v_y) are the components of \vec{v} . With this approach, all trajectories start with $\vec{v} = (0, 1)$.

Next, we normalize the local parameters, which denote the shape of the flagellum. We set $\mathbf{P}_{(0,i)}$ as the origin of the coordinate system. Then, we rotate the rest of the flagellum parameters ($\mathbf{P}_{(1,i)}, \mathbf{P}_{(2,i)}, \mathbf{P}_{(3,i)}, \mathbf{P}_{(4,i)}$) in each frame i an angle given by the respective orientation vector

$$\alpha_i = -\arctan\left(\frac{\alpha_x}{\alpha_y}\right)_i,$$

where i represent the trajectory index, and (α_x, α_y) are the components of $\vec{\alpha}$. This transformation is depicted in Fig. 1(a-c).

Diffusion model The Denoising Diffusion Probabilistic Model (DDPM) used in this study is implemented with a UNet [14] architecture. The model processes input data with dimensions $[B, T, \dim(\rho_t)]$, where B represents the batch size, T is the sequence length, and $\dim(\rho_t)$ denotes the dimensionality of the sperm model’s parameter vector. Specifically, as introduced in Section 4.2 The architecture is composed of three stages, starting with a downsampling phase consisting of four residual block [66] using 1D convolutions. Next, comes a middle phase comprising four residual blocks. Finally, the upsampling phase mirrors the downsampling, however, it uses transposed 1D convolutions. To enable the model to track the time step (H) of the denoising process, a sinusoidal positional encoding [63] followed by a few fully connected layers were used. The output of the time encoding layers is combined in each block of the UNet by addition. The model has 3.96M parameters, requiring 15MB of free memory to be allocated. In our experiments setup, the training procedure consumed 1.1GB of memory from a GPU NVIDIA RTX 2070. The hyperparameters used for training the model are detailed in Table 3.

CycleGAN In this work, a CycleGAN consisting of two generators and two discriminators is used to facilitate image-to-image translation without paired data. The generators are based on U-Net [14] architecture and employ Least Squares

GAN (LSGAN) [67] for loss function training. Each generator takes an input image of size $(500 \times 500 \times 3)$ and consists of three stages. The first stage is a convolutional downsampling phase. Next, few residual blocks consisting of two convolutional layers and skip connections are applied. Finally, the upsampling stage mirrors the downsampling, however, it uses transposed convolutions. This process generate an output of shape $(500 \times 500 \times 3)$.

The two discriminator models take input images of size $(500 \times 500 \times 3)$. Then convolutional layers are applied to downsample the image by a factor of 3. The final layer outputs a scalar value representing the likelihood that the input image is real or fake. The LSGAN loss function is used to train the discriminator.

Each generator has 11.4M parameters and requires 43 MB of memory. Each discriminator has 2.8M parameters and requires 10MB of memory. In our experiments, the training process consumed 6.25 GB of memory on a GPU NVIDIA RTX 2070, with 8 GB of GDDR6 memory.

LSTM The Long Short-Term Memory (LSTM) model used as baseline follows a common straightforward architecture for processing temporal sequences. The input to the model is a tensor with dimensions $[B, \iota, \dim(\rho_t)]$, where B represents the batch size, ι is the length of the input sequence set to $\iota = 4$ in our experiments, and $\dim(\rho_t)$ denotes the dimensionality of the sperm model’s parameter vector. The core of the model is a single LSTM layer, which captures temporal dependencies within the data. Then, the model includes two fully connected layers that process the output of the LSTM and map it to the output space of shape $[B, 1, \dim(\rho_t)]$. Section 4.2 provides further details on this model.

Gaussian model The Gaussian model is used as baseline and is introduced in Section 4.2. Here we provide the μ and Σ computed from the training sequences in Table 4. Note that since Σ is a diagonal matrix, we just represent μ and σ for each parameter of the spermatozoon model.

Table 3 Training hyperparameters for the DDPM model

Parameter	Value/Details
Denoising steps (H)	20
EMA decay	0.995
Input normalization	$\frac{x_i - \mu_i}{\sigma_i}$
Batch size (B)	64
Learning rate	2×10^{-5}
Sequence length (T)	16
$\dim(\rho_t)$	14

Table 4 Value of Mean (μ) and Standard Deviation (σ) of sperm model parameters

Parameter	Mean (μ)	Standard Deviation (σ)
\mathbf{P}_0	(−0.710, 0.012)	(0.186, 0.060)
\mathbf{P}_1	(−0.472, −0.011)	(0.168, 0.147)
\mathbf{P}_2	(−0.302, 0.013)	(0.130, 0.196)
\mathbf{P}_3	(−0.111, 0.036)	(0.158, 0.101)
\mathbf{P}_4	(0.111, −0.012)	(0.054, 0.036)
\vec{v}	(3.637, −0.000)	(1.313, 0.001)
$\vec{\alpha}$	(0.892, −0.045)	(0.336, 0.299)

Author Contributions S. Hernández-García and A. Cuesta-Infante are the main contributors for this paper, conceptualizing the key ideas, conducting the research, and writing the manuscript; A. Cuesta-Infante also provided guidance and supervision across the research; D. Makris contributed with ideas and supervision of the methodology, and revising the manuscript; and A. S. Montemayor supervised and participated in the manuscript revision.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work was supported by: R&D project TED2021-129162B-C22, funded by MICIU/AEI/10.13039/501100011033 and the European Union NextGenerationEU/ PRTR; and R&D project PID2021-128362OB-I00, funded by MICIU/AEI/10.13039/501100011033/ and FEDER/UE.

Data Availability Real videos can be provided upon request and for research purposes only. Synthetically generated videos are available at <https://github.com/SergioHdezG/sperm-diffuser>.

Declarations

Compliance with Ethical Standards The real videos of porcine semen used in this research were provided by an ISO (9001:2015) certified company.

Competing interests The authors have no commercial or competing interests to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Chawre S, Khatib MN, Rawekar A, Mahajan S, Jadhav R, More A (2024) A review of semen analysis: updates from the who sixth edition manual and advances in male fertility assessment. *Cureus* 16(6):63485
- Panner Selvam MK, Moharana AK, Baskaran S, Finelli R, Hudnall MC, Sikka SC (2024) Current updates on involvement of artificial intelligence and machine learning in semen analysis. *Medicina* 60(2):279
- Finelli R, Leisegang K, Tumallapalli S, Henkel R, Agarwal A (2021) The validity and reliability of computer-aided semen analyzers in performing semen analysis: a systematic review. *Transl Androl Urol* 10(7):3069
- Kwon W-S, Shin D-H, Ryu D-Y, Khatun A, Rahman MS, Pang M-G (2017) Applications of capacitation status for litter size enhancement in various pig breeds. *Asian-Australas J Anim Sci* 31(6):842
- van der Horst G (2020) Computer aided sperm analysis (casa) in domestic animals: current status, three d tracking and flagellar analysis. *Anim Reprod Sci* 220:106350. <https://doi.org/10.1016/j.anireprosci.2020.106350>
- O'meara C, Henrotte E, Kupisiewicz K, Latour C, Broekhuijse M, Camus A, Gavin-Plagne L, Sellem E (2022) The effect of adjusting settings within a computer-assisted sperm analysis (casa) system on bovine sperm motility and morphology results. *Anim Reprod* 19(1):20210077
- Organization WH et al (2010) Who laboratory manual for the examination and processing of human semen
- Tomlinson MJ (2016) Uncertainty of measurement and clinical value of semen analysis: has standardisation through professional guidelines helped or hindered progress? *Andrology* 4(5):763–770. <https://doi.org/10.1111/andr.12209>
- Gallagher MT, Smith D, Kirkman-Brown J (2018) Casa: tracking the past and plotting the future. *Reprod Fertil Dev* 30(6):867–874
- Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
- Hidayatullah P, Mengko T, Munir R (2017) A survey on multisperm tracking for sperm motility measurement. *Int J Mach Learn Comput* 7(5):144–151
- Zhang Z, Qi B, Ou S, Shi C (2022) Real-time sperm detection using lightweight yolov5. In: 2022 IEEE 8th International Conference on Computer and Communications (ICCC), pp 1829–1834. <https://doi.org/10.1109/ICCC56324.2022.10065602>
- Jabbari H, Bigdeli N (2023) New conditional generative adversarial capsule network for imbalanced classification of human sperm head images. *Neural Comput Appl* 1–16
- Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18. Springer, pp 234–241
- Marín R, Chang V (2021) Impact of transfer learning for human sperm segmentation using deep learning. *Comput Biol Med* 136:104687
- Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 10684–10695
- Yüzkat M, İlhan HO, Aydın N (2021) Multi-model cnn fusion for sperm morphology analysis. *Comput Biol Med* 137:104790. <https://doi.org/10.1016/j.compbiomed.2021.104790>
- Mashaal AA, Eldosoky MA, Mahdy LN, Kadry AE (2022) Automatic healthy sperm head detection using deep learning. *Int J Adv Comput Sci Appl* 13(4)
- Hidayatullah P, Wang X, Yamasaki T, Mengko TLER, Munir R, Barlian A, Sukmawati E, Suprptono S (2021) Deep sperm: A robust and real-time bull sperm-cell detection in densely populated semen videos. *Comput Methods Prog Biomed* 209:106302. <https://doi.org/10.1016/j.cmpb.2021.106302>
- Arasteh A, Vosoughi Vahdat B, Salman Yazdi R (2018) Multi-target tracking of human spermatozoa in phase-contrast microscopy image sequences using a hybrid dynamic bayesian network. *Sci Rep* 8(1):5068
- Xu F, Guo G, Zhu W, Fan L (2018) Human sperm acrosome function assays are predictive of fertilization rate in vitro: a retrospective cohort study and meta-analysis. *Reprod Biol Endocrinol* 16:1–29
- Li C, Ni Y, Yao L, Fang J, Jiang N, Chen J, Lin W, Ni H, Zheng H (2024) The correlation between sperm percentage with a small acrosome and unexplained in vitro fertilization failure. *BMC Pregnancy Childbirth* 24(1):58
- Hernandez-Herrera P, Montoya F, Rendón-Mancha JM, Darszon A, Corkidi G (2018) 3-d +t human sperm flagellum tracing in low snr fluorescence images. *IEEE Trans Med Imaging* 37(10):2236–2247. <https://doi.org/10.1109/TMI.2018.2840047>

24. Battistella A, Andolfi L, Stebel M, Ciubotaru C, Lazzarino M (2023) Investigation on the change of spermatozoa flagellar beating forces before and after capacitation. *Biomater Adv* 145:213242. <https://doi.org/10.1016/j.bioadv.2022.213242>
25. Hyakutake T, Suzuki H, Yamamoto S (2015) Effect of non-newtonian fluid properties on bovine sperm motility. *J Biomech* 48(12):2941–2947
26. Takei GL, Fujinoki M, Yoshida K, Ishijima S (2017) Regulatory mechanisms of sperm flagellar motility by metachronal and synchronous sliding of doublet microtubules. *MHR: Basic Science of Reproductive Medicine* 23(12):817–826
27. Bukatin A, Kukhtevich I, Stoop N, Dunkel J, Kantsler V (2015) Bimodal rheotactic behavior reflects flagellar beat asymmetry in human sperm cells. *Proc Natl Acad Sci* 112(52):15904–15909
28. Oriola D, Gadêlha H, Casademunt J (2017) Nonlinear amplitude dynamics in flagellar beating. *R Soc Open Sci* 4(3):160698
29. Gallagher MT, Cupples G, Ooi EH, Kirkman-Brown J, Smith D (2019) Rapid sperm capture: high-throughput flagellar waveform analysis. *Hum Reprod* 34(7):1173–1185
30. İlhan HO, Sigirci IO, Serbes G, Aydin N (2020) A fully automated hybrid human sperm detection and classification system based on mobile-net and the performance comparison with conventional methods. *Med Biol Eng Comput* 58:1047–1068
31. Chang V, Saavedra JM, Castañeda V, Sarabia L, Hitschfeld N, Härtel S (2014) Gold-standard and improved framework for sperm head segmentation. *Comput Methods Programs Biomed* 117(2):225–237
32. İlhan HO, Yuzkat M, Aydin N (2021) Sperm motility analysis by using recursive kalman filters with the smartphone based data acquisition and reporting approach. *Expert Syst Appl* 186:115774. <https://doi.org/10.1016/j.eswa.2021.115774>
33. Ottl S, Amiriparian S, Gerczuk M, Schuller BW (2022) motilitai: A machine learning framework for automatic prediction of human sperm motility. *Iscience* 25(8)
34. Shaker F, Monadjemi SA, Alirezaie J, Naghsh-Nilchi AR (2017) A dictionary learning approach for human sperm heads classification. *Comput Biol Med* 91:181–190. <https://doi.org/10.1016/j.combiomed.2017.10.009>
35. Haugen TB, Hicks SA, Andersen JM, Witczak O, Hammer HL, Borgli R, Halvorsen P, Riegler M (2019) Visem: A multimodal video dataset of human spermatozoa. In: *Proceedings of the 10th ACM Multimedia Systems Conference. MMSys '19*, pp 261–266. <https://doi.org/10.1145/3304109.3325814>
36. Thambawita V, Hicks SA, Storås AM, Nguyen T, Andersen JM, Witczak O, Haugen TB, Hammer HL, Halvorsen P, Riegler MA (2023) Visem-tracking, a human spermatozoa tracking dataset. *Sci Data* 10(1):1–8
37. Chen A, Li C, Zou S, Rahaman MM, Yao Y, Chen H, Yang H, Zhao P, Hu W, Liu W, Grzegorzeczek M (2022) Svia dataset: A new dataset of microscopic videos and images for computer-aided sperm analysis. *Biocybern Biomed Eng* 42(1):204–214
38. Yuzkat M, İlhan HO, Aydin N (2023) Detection of sperm cells by single-stage and two-stage deep object detectors. *Biomed Signal Process Control* 83:104630. <https://doi.org/10.1016/j.bspc.2023.104630>
39. Arasteh A, Vahdat BV (2016) Evaluation of multi-target human sperm tracking algorithms in synthesized dataset. *Int J Monit Surveillance Technol Res* 4(2):16–29
40. Choi J-W, Alkhoury L, Urbano LF, Masson P, VerMilyea M, Kam M (2022) An assessment tool for computer-assisted semen analysis (casa) algorithms. *Sci Rep* 12(1):16830
41. Hernández-Ferrándiz D, Pantrigo JJ, Cabido R (2022) Scasa: From synthetic to real computer-aided sperm analysis. In: *Bio-inspired systems and applications: from robotics to ambient intelligence*. Springer, pp 233–242
42. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. *Adv Neural Inf Process Syst* 27:1
43. Balayev K, Guluzade N, Aygün S, İlhan HO (2021) The implementation of dcgan in the data augmentation for the sperm morphology datasets. *Avrupa Bilim ve Teknoloji Dergisi*. (26), 307–314
44. Paul D, Tewari A, Jeong J, Banerjee I (2021) Boosting classification accuracy of fertile sperm cell images leveraging cdcgan. *ICLR*
45. Abbasi A, Bahrami S, Hemmati T, Mirroshandel SA (2023) Transfer-gan: data augmentation using a fine-tuned gan for sperm morphology classification. *Comput Methods Biomech Biomed Eng Imaging Vis* 0(0):1–17
46. Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. *Adv Neural Inf Process Syst* 33:6840–6851
47. Kingma DP, Welling M (2014) Auto-encoding variational bayes. In: *ICLR*
48. Van Den Oord A, Kalchbrenner N, Kavukcuoglu K (2016) Pixel recurrent neural networks. In: *International conference on Machine Learning*. PMLR, pp 1747–1756
49. Nichol AQ, Dhariwal P (2021) Improved denoising diffusion probabilistic models. In: *Proceedings of the 38th International conference on machine learning*, pp 8162–8171
50. Saharia C, Chan W, Chang H, Lee C, Ho J, Salimans T, Fleet D, Norouzi M (2022) Palette: Image-to-image diffusion models. In: *ACM SIGGRAPH 2022*, pp 1–10
51. Luo S, Hu W (2021) Diffusion probabilistic models for 3d point cloud generation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 2837–2845
52. Austin J, Johnson DD, Ho J, Tarlow D, Van Den Berg R (2021) Structured denoising diffusion models in discrete state-spaces. *Adv Neural Inf Process Syst* 34:17981–17993
53. Rasul K, Seward C, Schuster I, Vollgraf R (2021) Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In: *International conference on machine learning*. PMLR, pp 8857–8868
54. Yang L, Zhang Z, Song Y, Hong S, Xu R, Zhao Y, Zhang W, Cui B, Yang M-H (2022) Diffusion models: a comprehensive survey of methods and applications. *ACM Comput Surv*
55. Janner M, Du Y, Tenenbaum JB, Levine S (2022) Planning with diffusion for flexible behavior synthesis. In: *International conference on machine learning*
56. Jiang CM, Cormman A, Park C, Sapp B, Zhou Y, Anguelov D (2023) Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp 9644–9653
57. Tseng J, Castellon R, Liu K (2023) Edge: Editable dance generation from music. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp 448–458
58. Huang S, Wang Z, Li P, Jia B, Liu T, Zhu Y, Liang W, Zhu S-C (2023) Diffusion-based generation, optimization, and planning in 3d scenes. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 16750–16761
59. Rempe D, Luo Z, Bin Peng X, Yuan Y, Kitani K, Kreis K, Fidler S, Litany O (2023) Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp 13756–13766
60. Fitzgibbon AW, Fisher RB (1995) A buyer's guide to conic fitting. In: *Proceedings of the 6th British conference on machine vision (Vol. 2)*. BMVC '95, pp 513–522. BMVA Press
61. Zhu J-Y, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE International conference on computer vision*, pp 2223–2232
62. Hernández-Ferrándiz D, Oliva-Wilkinson I, Pantrigo JJ, Cabido R (2023) Recasa: real-time computer-assisted sperm analysis. In:

- Real-time processing of image, depth and video information 2023, vol 12571. SPIE, pp 22–30
63. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser LU, Polosukhin I (2017) Attention is all you need. In: *Advances in Neural Information Processing Systems*, vol 30
 64. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X (2016) Improved techniques for training gans. *Adv Neural Inf Process Syst* 29
 65. Hochreiter S, Schmidhuber J (1997) long short-term memory. *Neural Comput* 9(8):1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
 66. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 770–778
 67. Mao X, Li Q, Xie H, Lau RY, Wang Z, Paul Smolley S (2017) Least squares generative adversarial networks. In: *Proceedings of the IEEE international conference on computer vision*, pp 2794–2802

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.