Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Bayesian Mixture-of-Views Framework for Joint Segmentation and Focal-Depth Prediction in Multi-Focal Oocyte Microscopy

Godwin Komla Krampah
Vrije Universiteit Brussel
Samuel Wauthier
Vrije Universiteit Brussel
Ann Dooms
Vrije Universiteit Brussel
Koen Wouters
Brussels IVF, Center for Reproductive Medicine
Michel De Vos
Brussels IVF, Center for Reproductive Medicine
Tan Lu
tan.lu@vub.be

Vrije Universiteit Brussel

Article

Keywords:

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11366021/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Bayesian Mixture-of-Views Framework for Joint Segmentation and Focal-Depth Prediction in Multi-Focal Oocyte Microscopy

Godwin Komla Krampah[1], Samuel Wauthier[1], Ann Dooms[1], Koen Wouters[2], Michel De Vos[2,3], and Tan Lu[1,4,*]

[1]Department of Mathematics and Data Science, Vrije Universiteit Brussel, 1050 Brussels, Belgium
[2]Brussels IVF, Center for Reproductive Medicine, UZ Brussel, 1090 Brussels, Belgium
[3]Research Group Genetics, Reproduction and Development (GRAD), Vrije Universiteit Brussel, 1050 Brussels, Belgium
[4]Data Science Lab, Department of Digitization, Royal Library of Belgium (KBR), 1000 Brussels, Belgium
[*]tan.lu@vub.be

ABSTRACT

Human oocyte structures are not all clearly visible at a single focal depth, so embryologists scan several planes, yet published segmentation methods analyse one. We present a Bayesian mixture-of-views (BMoV) framework that segments the cell body, perivitelline space (PVS), polar body (PB) and refractile body (RB) from an 11-plane focal stack (-75 to $+75\,\mu$m) and predicts, for each structure in each oocyte, the depths at which it is clearest. Of 897 oocytes, 119 were held out for evaluation. On this test set, Dice was 0.969 for the cell body, 0.790 for the PVS, 0.774 for the PB and 0.218 for the RB. The predicted depth profiles ranked annotated depths well above chance (precision–recall AUC 0.79), and the two highest-ranked depths included an annotated one in 77% of cases. These profiles come at no cost: segmentation was as accurate as when all depths were weighted equally. Using all depths rather than one improved segmentation most for the smallest structures, and self-supervised pretraining on unannotated oocyte images was the most beneficial design choice. The framework can point embryologists to the depths worth examining. Refractile-body segmentation, limited by its small size and rarity, is the main target for future work.

Introduction

During in vitro fertilisation (IVF), the morphological assessment of human oocytes has been proposed to inform oocyte selection, although its predictive value for most individual features is limited[1,2]. Selection matters when only a subset of the retrieved oocytes will be fertilised or vitrified. Because oocytes are three-dimensional biological objects, the semantic structures relevant to this assessment - the cell body, perivitelline space (PVS), polar body (PB), and refractile body (RB) - are not uniformly visible at a single focal depth: each structure, and even the same structure across different samples, may be clearly resolved only within a subset of the focal-depth range captured during multi-focal microscopy[3]. Fixed single-plane imaging therefore cannot capture all structures with consistent clarity. Current clinical practice relies on embryologists manually scanning through multiple focal planes and visually assessing structure boundaries, a process that introduces subjectivity and inter-observer variability, limiting reproducibility and standardisation of oocyte assessment[1,2]. This limitation is particularly consequential for PB and RB, whose presence and morphology have been associated with blastocyst development and implantation outcomes[4,5].

Convolutional neural networks (CNNs) have shown promising performance in segmenting oocytes and other biological structures[6], but published approaches to oocyte segmentation analyse a single selected focal plane and do not model the focal-depth axis explicitly[7,8]. Structural features that are well defined at one depth may become blurred or absent at another, whereas embryologists dynamically adjust the microscope's focus to examine each structure at the depth or depths where it is most clearly visible - a behaviour current automated oocyte segmentation systems do not emulate.

We propose a probabilistic deep-learning framework for oocyte image analysis that unifies semantic structure segmentation with the prediction of each structure's visibility across the focal-depth stack, through an *amortised Bayesian Mixture-of-Views (BMoV)* formulation. Rather than treating each focal plane as visually independent, BMoV interprets multi-focal imaging as a probabilistic mixture over focal depths, decomposing pixel-level structure classification into depth-conditioned components that are combined according to a learned, per-sample, per-structure focal-plane admissibility distribution. Amortised inference[9] allows this to be learned efficiently via shared latent representations, integrating three components, formalised in Methods (equation (6)): (i) depth-conditioned segmentation posterior estimation, (ii) per-structure focal-plane visibility prediction, and

(iii) probabilistic aggregation across the focal-depth stack.

This study examines whether such a model can jointly produce a composite segmentation of four oocyte structures and predict structure-specific focal-depth profiles from an 11-plane image stack (-75 to $+75$ μm). The framework was developed on 778 annotated oocytes, using controlled cross-validation comparisons to examine focal coverage, aggregation across depths and model initialisation, and evaluated on an independent, patient-disjoint set of 119. To our knowledge, this is among the first frameworks to perform semantic structure segmentation and per-structure focal-plane visibility prediction jointly for multi-focal oocyte microscopy, reflecting the multi-plane assessment strategy used in embryology practice.

Related work

Deep learning has substantially advanced the automatic delineation of structures in microscopy images, with architectures such as U-Net[10], DeepLab[11,12], and Cellpose[13] now widely adopted in biomedical microscopy[14]. Volumetric variants of these architectures exist, but the oocyte segmentation literature built on them reports a single segmentation without indicating the depths at which each structure is assessable. We instead build on probabilistic mixture modelling[15] to represent the fact that different oocyte structures, and the same structure across different samples, may be clearly visible only at certain focal depths.

In biological microscopy, segmentation has commonly been applied to single images captured at one focal depth. Cohen and Uhlmann[16] segmented phase-contrast microscopy images of smooth muscle cells using a U-Net-based method. Specifically for oocyte images, Firuzinia et al.[3] developed a robust multiclass segmentation method for human metaphase II oocytes. Targosz et al.[6] compared several segmentation architectures on 334 oocyte images annotated with 13 morphological classes and obtained a weighted intersection-over-union (IoU — the overlap between the predicted and annotated regions, on a scale from 0 to 1) of 0.897 with a DeepLab-v3 model; in later work[17] they used DeepLabV3+ to segment 766 oocyte images (test IoU 0.66) as a step towards classifying oocyte maturity stage, which they achieved with 95.7% test accuracy. Fjeldstad et al.[8] reported IoU of 97–98% with a proprietary transformer-based model applied to single-plane images, for a different set of structures (ooplasm, PVS and zona pellucida). These values are not directly comparable with those reported here: each was obtained against an annotation drawn on the same single plane the model received, and on different target structures.

To address the limitation of single-plane input, multi-focal image stacks have been used in other microscopy settings, including pollen recognition[18] and malaria diagnosis from blood smears[19]. Within reproductive medicine, Souda et al.[20] compared models using a single focal plane against models using seven focal planes for embryo stage classification, and found the multi-focal models performed better (F1-score - the harmonic mean of precision and recall - of 0.87 against 0.80). Jonaitis et al.[21] fused multiple focal planes with a transformer-based model and raised accuracy in classifying embryo developmental stage from 84.7% to 94.1% relative to single-plane images. These approaches exploit the three-dimensional focal-stack structure, but none indicates at which focal depths individual structures are visible.

More broadly, multi-focal microscopy images are often combined using image fusion[22,23] or focus stacking[24], techniques that compress the focal stack into a single all-in-focus composite image; this convenience can come at the cost of depth-specific structural information. BMoV instead predicts each structure's distribution over annotated focal depths directly, preserving this information and providing a per-structure view of the depths at which each structure was assessed.

Methods

Bayesian Mixture-of-Views (BMoV)

Given a set of class labels $\mathscr{S}$ with cardinality $|\mathscr{S}| = C$, comprising the four foreground structures of interest (background is handled separately, as described in equation (7)), the semantic structure segmentation task can be characterised by a binary random variable $T_s(u)$ defined for each structure $s \in \mathscr{S}$ and for each pixel u as:

$$T_s(u) = \begin{cases} 1, & \text{if pixel } u \text{ belongs to structure } s, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The posterior probability $\mathbb{P}(T_s(u) = 1 \mid \mathbf{X})$ is considered, where $\mathbf{X}$ denotes a comprehensive visual representation of an oocyte sample, here the multi-focal acquisition described in Dataset and annotation.

Unlike conventional 2D image processing, the focal depth at which an oocyte structure is clearly visible varies across structures and even across samples for the same structure. To model this variability explicitly, a continuous latent focal-view variable is introduced for each structure s,

$$Z_s \in [-F, F], \quad F \in \mathbb{R},\ F > 0,$$

where F denotes the positive half-width of the sampled focal-depth range ($F = 75$ μm for the acquisitions used here), with conditional probability density $p_s(z \mid \mathbf{X})$. That is, Z_s represents an admissible view (e.g. focal depth) at which structure s may

be observed. For readability, the subscript s is suppressed on the density below wherever no ambiguity arises. The posterior is expressed via a mixture-of-views (MoV) expectation over Z_s:

$$\begin{aligned}&\mathbb{P}(T_s(u)=1 \mid \mathbf{X})\\&\quad=\int_{-F}^{F}\mathbb{P}(T_s(u)=1 \mid Z_s=z,\mathbf{X})\,p(z \mid \mathbf{X})\,\mathrm{d}z\,.\end{aligned}\tag{2}$$

The acquired focal planes are physically discrete, so equation (2) is a conceptual device rather than a quantity that is evaluated: the integral is approximated by a finite mixture over a partition of $[-F,F]$ into K measurable intervals $\{B_k\}_{k=1}^{K}$, each represented by a single depth. Writing $\mathbb{P}(z_s=k \mid \mathbf{X})=\int_{B_k}p(z \mid \mathbf{X})\,\mathrm{d}z$ for the probability mass assigned to interval B_k, and introducing a latent discrete random variable z_s supported on $\{1,2,\ldots,K\}$, this gives:

$$\begin{aligned}&\mathbb{P}(T_s(u)=1 \mid \mathbf{X})=\\&\quad\sum_{k=1}^{K}\mathbb{P}(T_s(u)=1 \mid z_s=k,\mathbf{X})\,\mathbb{P}(z_s=k \mid \mathbf{X})\,.\end{aligned}\tag{3}$$

Equation (3) is exact for a categorical view variable; approximating its components by values at representative depths is a separate step, whose discretisation error is bounded under Lipschitz continuity of the view-conditioned posterior (Supplementary Note S1). That bound concerns discretisation of a fixed continuous model only: it does not bound the estimation error of the learned network, establish calibration of the predicted distributions, or establish that the 11 acquired planes suffice for every structure.

The second term, $\mathbb{P}(z_s=k \mid \mathbf{X})$, represents the probability that, for a particular sample $\mathbf{X}$, structure s is annotated as most clearly visible at focal plane k. This probability varies from sample to sample and depends on the positioning of individual oocytes during imaging: the view or views through which the same structure (e.g. the PB) is visible may differ across samples due to variation in orientation and positioning. Capturing this conditional probability explicitly is therefore necessary, and it cannot be represented by a single fixed focal plane.

The first term, $\mathbb{P}(T_s(u)=1 \mid z_s=k,\mathbf{X})$, denotes the probability that, given the visual representation $\mathbf{X}$ and conditioning on view k for structure s, a pixel u belongs to structure s. Here $g_k(\mathbf{X})$ is taken as an approximating, plane-specific representation such that

$$\mathbb{P}(T_s(u)=1 \mid z_s=k,\mathbf{X})\approx\mathbb{P}(T_s(u)=1 \mid z_s=k,g_k(\mathbf{X}))\,,\tag{4}$$

where $g_k(\mathbf{X})$ denotes the representation associated with view k. In the implementation, the encoder consumes the complete 11-plane stack, so $g_k(\mathbf{X})$ denotes the k-th view-conditioned output of the decoder computed from this shared representation, rather than an observation restricted to plane k. Equation (4) should therefore be read as indexing the view rather than as discarding the remaining planes (see Implementation details).

Applying Bayes' rule to this probability gives:

$$\begin{aligned}&\mathbb{P}(T_s(u)=1 \mid z_s=k,g_k(\mathbf{X}))=\\&\frac{p(g_k(\mathbf{X}) \mid z_s=k,T_s(u)=1)\,\mathbb{P}(T_s(u)=1 \mid z_s=k)}{\sum_{l\in\{0,1\}}p(g_k(\mathbf{X}) \mid z_s=k,T_s(u)=l)\,\mathbb{P}(T_s(u)=l \mid z_s=k)}\,,\end{aligned}$$

which simplifies to

$$\frac{\lambda_k^{+}\pi_k^{+}}{\lambda_k^{+}\pi_k^{+}+\lambda_k^{-}\pi_k^{-}}\,.$$

That is, Bayesian modelling reformulates this posterior as a ratio between positive and negative likelihoods and priors, where $\lambda_k^{\pm}$ and $\pi_k^{\pm}$ denote the *view-conditioned likelihoods* and the corresponding *prior probabilities*, respectively, defined as:

$$\begin{aligned}\lambda_k^{\pm}&=p\big(g_k(\mathbf{X})\,\big|\,z_s=k,T_s(u)=\tfrac{1}{2}\pm\tfrac{1}{2}\big)\,,\\\pi_k^{\pm}&=\mathbb{P}\big(T_s(u)=\tfrac{1}{2}\pm\tfrac{1}{2}\,\big|\,z_s=k\big)\,.\end{aligned}$$

The superscript $+$ corresponds to $T_s(u)=1$ (pixel u belongs to structure s) and the superscript $-$ to $T_s(u)=0$, so that $\lambda_k^{\pm}$ are the view-conditioned likelihoods of the observation under the two hypotheses and $\pi_k^{\pm}$ the corresponding prior probabilities. For

readability, the dependence of these likelihood and prior terms on structure s, pixel location u, and the input $\mathbf{X}$ is suppressed in the notation above and throughout the remainder of this subsection.

This ratio can be parameterised through neural network logits, such that:

$$\begin{aligned}
&\mathbb{P}(T_s(u) = 1 \mid z_s = k, g_k(\mathbf{X})) \\
&\quad = \frac{\lambda_k^+ \pi_k^+}{\lambda_k^+ \pi_k^+ + \lambda_k^- \pi_k^-} \\
&\quad = \sigma\left(\log \frac{\lambda_k^+}{\lambda_k^-} + \log \frac{\pi_k^+}{\pi_k^-} \right),
\end{aligned} \tag{5}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $\sigma(x) = \frac{1}{1+e^{-x}}$.

Consequently, the MoV model (equation (3)) with Bayesian expansion can be *amortised* using neural networks as:

$$\begin{aligned}
&\mathbb{P}(T_s(u) = 1 \mid \mathbf{X}) \\
&\quad = \sum_{k=1}^{K} \mathbb{P}(T_s(u) = 1 \mid z_s = k, \mathbf{X}) \, \mathbb{P}(z_s = k \mid \mathbf{X}) \\
&\quad = \sum_{k=1}^{K} \sigma\left(\log \frac{\lambda_k^+}{\lambda_k^-} + \log \frac{\pi_k^+}{\pi_k^-} \right) \mathbb{P}(z_s = k \mid \mathbf{X}) \, .
\end{aligned} \tag{6}$$

This framework consists of three components: (i) the view-conditioned posterior $\mathbb{P}(T_s(u) = 1 \mid z_s = k, \mathbf{X})$, approximated through an encoder–decoder network that outputs $\sigma\big(\log \frac{\lambda_k^+}{\lambda_k^-} + \log \frac{\pi_k^+}{\pi_k^-}\big)$; (ii) a per-structure focal-plane distribution $\mathbb{P}(z_s = k \mid \mathbf{X})$, modelled via a view-prediction network; and (iii) an across-view aggregation module, i.e. $\sum_{k=1}^{K}(\cdot)$, that marginalises the product of the per-view posterior and the visibility distribution to produce the unified per-pixel segmentation posterior $\mathbb{P}(T_s(u) = 1 \mid \mathbf{X})$ over all discretised view conditions.

The Bayesian expansion in equation (5) motivates the form of the view-conditioned posterior. In the implementation, the posterior logits and the focal-depth weights are learned directly by a supervised neural network: the likelihood and prior terms are not estimated separately, no distribution is placed over the network parameters, and the calibration of the resulting output probabilities was not evaluated. The framework is Bayesian in the sense of amortised point estimation of a mixture posterior, and no uncertainty quantification over parameters is claimed.

Amortised neural network for BMoV

Amortised learning Given an input oocyte sample $\mathbf{X}$ acquired through multi-focal imaging, a shared encoder backbone (Fig. 1a) extracts view-sensitive representations that enable amortised learning of both the focal-plane distribution and view-aware segmentation, via the *view head* and the *segmentation head* respectively.

View head The view head, implemented as a lightweight fully connected module attached to the encoder bottleneck, learns the per-structure focal-plane distribution $\mathbb{P}(z_s = k \mid \mathbf{X})$ (Fig. 1a).

View-dependent segmentation head The view-dependent segmentation head is a U-Net-style decoder with skip connections, attached to the shared encoder backbone, and estimates the per-structure, per-view segmentation posterior $\mathbb{P}(T_s(u) = 1 \mid z_s = k, \mathbf{X})$, parameterised through neural-network logits as in equation (5) (Fig. 1b).

Per-structure segmentation via MoV The final per-structure segmentation posterior, $\mathbb{P}(T_s(u) = 1 \mid \mathbf{X})$, is obtained by marginalising the discrete view variable: per-view posteriors $\mathbb{P}(T_s(u) = 1 \mid z_s = k, \mathbf{X})$ are aggregated, weighted by the learned focal-plane distribution $\mathbb{P}(z_s = k \mid \mathbf{X})$, allowing the network to infer a unified segmentation belief by integrating evidence across focal depths (Fig. 1c).

Cascading normalisation for hierarchical plane stitching The resulting probabilities $P_s \triangleq \mathbb{P}(T_s(u) = 1 \mid \mathbf{X})$ are defined independently for each semantic structure s and are therefore not normalised across class labels. Clinical assessment and downstream analysis, however, generally require a single priority-resolved composite segmentation rather than independent per-structure predictions.

To construct such a segmentation, Prioritised Cascading Normalisation (PCN) enforces a hierarchical selection mechanism using a predefined overlap-resolution order — the refractile body (RB), then the polar body (PB), then the perivitelline space (PVS), then the cell body — so that smaller, anatomically contained structures take precedence where predictions overlap. Let

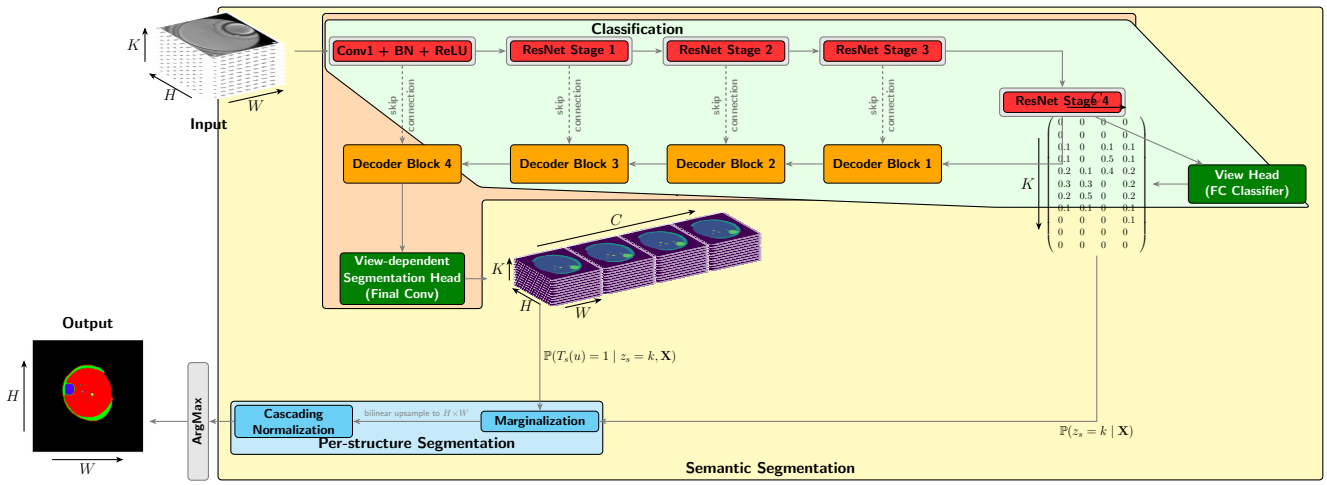

Figure 1. Amortised Bayesian Mixture-of-Views (BMoV) framework. **a**, A shared ResNet encoder extracts multi-scale, view-sensitive representations from the multi-focal input $\mathbf{X}$, supporting amortised learning across viewing conditions. The *view head* (a lightweight fully connected module) learns the per-structure focal-plane distribution $\mathbb{P}(z_s = k \mid \mathbf{X})$. **b**, The *view-dependent segmentation head*, implemented as a U-Net-style decoder with skip connections, estimates per-structure, per-view posteriors $\mathbb{P}(T_s(u) = 1 \mid z_s = k, \mathbf{X})$. **c**, Probabilistic MoV aggregation marginalises the discrete view variable to produce the unified per-structure posterior $\mathbb{P}(T_s(u) = 1 \mid \mathbf{X})$. Prioritised Cascading Normalisation (PCN) then stitches structure-specific predictions into a single priority-resolved 2D composite segmentation by enforcing a hierarchical semantic ordering; the exclusive label map is obtained by taking the arg max over the PCN categorical distribution $\{Q_{\text{bg}}, Q_{s_1}, \ldots, Q_{s_C}\}$ at each pixel. The numeric matrix shown is an illustrative $\mathbb{P}(z_s = k \mid \mathbf{X})$ and does not correspond to a particular sample.

the semantic labels be ordered by this priority, $\{s_1, s_2, \ldots, s_C\}$ with $s_1 = \text{RB}$ the highest-priority structure. PCN converts the independent probabilities P_{s_i} into a priority-resolved categorical distribution:

$$Q_{s_i} = P_{s_i} \prod_{j=1}^{i-1} \left(1 - P_{s_j}\right), \qquad Q_{\text{bg}} = \prod_{j=1}^{C} \left(1 - P_{s_j}\right), \tag{7}$$

where Q_{bg} denotes the background probability. By construction, $Q_{\text{bg}} + \sum_{i=1}^{C} Q_{s_i} = 1$, yielding a normalised segmentation map on a shared 2D plane. This operation "stitches" structure-specific predictions, each inferred while accounting for that structure's focal-plane distribution, into a single priority-resolved composite segmentation in which higher-priority structures dominate in regions of overlap.

Dataset and annotation

Cohort and imaging Multi-focal images were acquired with EmbryoScope during routine IVF treatment at Brussels IVF, UZ Brussel, between 2019 and 2023. Oocytes were imaged before intracytoplasmic sperm injection (ICSI). Of the approximately 18,000 oocytes imaged during this period, 897 were retained for analysis: those for which a complete set of expert annotation masks was available and whose acquisition passed the laboratory's image-quality check. Approximately 17,000 unannotated acquisitions from the same archive of the 18,000 oocytes, sharing no oocyte with the 897, formed the pool used for self-supervised pretraining (see "Data splits"). Each oocyte sample was imaged at 11 focal depths spanning -75 μm to $+75$ μm in increments of 15 μm. At each focal depth, the microscopy system acquires a grayscale image of size 800×800, yielding for each oocyte a 3D tensor of size $11 \times 800 \times 800$. In BMoV, this is the tensor $\mathbf{X}$ representing a multi-view observation of an oocyte across discretised focal depths, with the discrete view variable $z_s \in \{1, \ldots, 11\}$ indexing the depths $\{-75, -60, \ldots, +75\}$ μm in order.

Each stack was standardised plane by plane, each focal plane being centred and scaled by its own mean and standard deviation (Supplementary Table S6), and the four corner regions — which carry a burned-in acquisition label — were replaced before analysis (see "Augmentation and corner-label removal"). The normalisation constants vary systematically with depth, the plane mean rising from 0.504 at -75 μm to 0.588 at $+75$ μm and the standard deviation from 0.201 to 0.306, so the planes of a stack are not interchangeable in intensity or contrast. No other preprocessing was applied, and no registration was performed across planes.

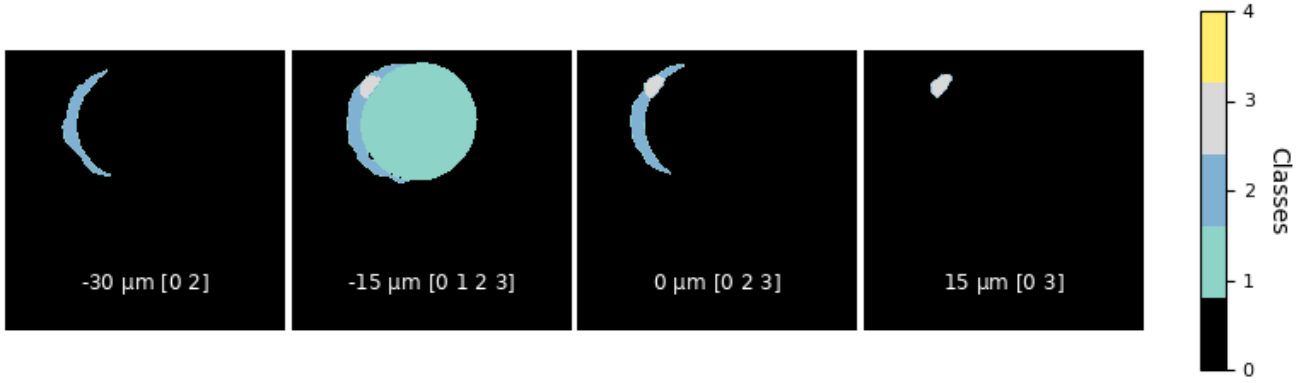

Figure 2. Annotation of semantic structures in an oocyte sample. Structures are annotated only at the focal depth or depths judged to provide the clearest view, illustrating structure-dependent visibility. Labels: background (0), cell body (1), PVS (2), PB (3), RB (4). Focal depths are given in μm.

Ethics approval and consent The research use of these routinely collected clinical data was approved by the Committee for Medical Ethics of UZ Brussel (Universitair Ziekenhuis Brussel; approval number 1432020000345) within the FROSCT project. The need for informed consent was waived by the Committee for Medical Ethics of UZ Brussel (Universitair Ziekenhuis Brussel) because the study used retrospectively collected, de-identified images from routine clinical care. All methods were carried out in accordance with relevant guidelines and regulations.

Data splits The development and test sets, and the ten cross-validation splits used for the design-choice experiments, were constructed at the level of the patient: all oocytes retrieved from a given patient were assigned to the same side of every split, so that no sibling oocyte appears in both development and test data. The development set comprises 778 oocytes from 187 patients; the test set comprises 119 oocytes from 21 patients. The split was stratified by the oocyte's morphological quality grade (four categories), and the grade distributions of the two sets did not differ significantly ($\chi^2 = 4.31$, 3 degrees of freedom, $P = 0.23$; Supplementary Table S5). Self-supervised pretraining (Implementation details) used the separate, unannotated pool of 17,000 oocyte acquisitions described above; this pool is disjoint from both the development and the test set, and no image from the 119 test oocytes was used at any stage of pretraining or training.

Quantities estimated from data and used during training were computed without reference to the test set: the inverse-frequency class weights were recomputed within each cross-validation fold on that fold's training split, and on the full development set for the final model.

Semantic structures Four semantic structures with differing spatial scales are considered: cell body (large), PVS (large–medium), PB (medium–small) and RB (small), annotated in 894, 862, 858 and 266 oocytes respectively across the full cohort of 897, before the development/test split. These four structures are referred to as classes 1–4 throughout. Per-structure counts within the development and test sets separately are given in Supplementary Table S5.

The objective of BMoV is twofold: (i) to infer, for every oocyte sample, the focal depths at which each semantic structure was judged to provide the clearest view, by estimating the focal-plane distribution $\mathbb{P}(z_s = k \mid \mathbf{X})$; and (ii) to produce an integrated segmentation in which each structure is inferred through expectation over all views (equation (3)). This inference is amortised at the sample level through joint learning of per-structure, per-view segmentation posteriors and per-structure focal-plane distributions, followed by PCN to stitch structure-specific predictions into a single priority-resolved composite segmentation.

Annotation procedure Annotation proceeded in two stages. First, as illustrated in Fig. 2, each semantic structure was independently annotated on the focal depth or depths selected as providing the clearest view, yielding structure-specific annotations associated with their class indices. Second, an aggregated mask was constructed by hierarchically fusing these annotations according to the semantic priority ordering of equation (7), producing a unified 2D representation. This aggregated mask serves as the ground truth for the single segmentation plane predicted by BMoV. An example aggregated mask, together with that sample's pixel distribution and focal-plane annotation, is shown in Fig. 3.

Because annotation was performed only on selected planes, the resulting focal-depth target represents the distribution of *selected* annotated views for a structure, and not an exhaustive record of every depth at which that structure is visible. A structure with no annotation in a given oocyte appears as background in that oocyte's segmentation target and is excluded from its focal-depth target, contributing no term to the focal-depth loss; the absence of an annotation does not by itself establish that

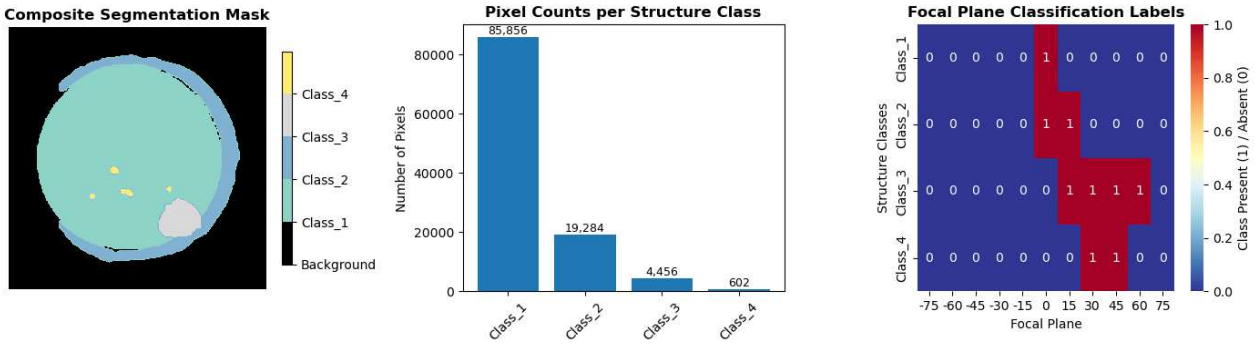

Figure 3. Aggregated segmentation target and annotation statistics for one example oocyte, the same sample shown in Fig. 2. **a**, Hierarchically aggregated joint segmentation target preserving overlap-resolution priority (RB, PB, PVS, then cell body, highest to lowest). **b**, Pixel counts for the four semantic structure classes in this example — cell body, PVS, PB and RB — illustrating their relative sizes within this single annotated sample; dataset-wide pixel-area statistics are given in Supplementary Table S1. **c**, This sample's per-structure focal-plane annotation, with red filling indicating the depths at which a structure was annotated as clearest; the corresponding dataset-wide annotation pattern is shown in Fig. 4.

the structure was not present.

A view-pattern analysis was conducted on the embryologists' annotations to characterise how different structures manifest across focal depths (Fig. 4), revealing both structure-specific and sample-to-sample variability. Fig. 4 shows that different semantic structures are annotated at different focal depths, and that even for the same structure the annotated depths vary across samples. Classes 1 and 2 show relatively sharp focal localisation, whereas Classes 3 and 4 show broader patterns: for the PB (Class 3), annotations occurred at every sampled depth across the cohort. This is the variability the proposed framework addresses: the focal-plane distribution $\mathbb{P}(z_s = k \mid \mathbf{X})$ is both structure- and sample-dependent, so no single preset focal plane can capture all semantic structures, or represent the same structure consistently across oocytes.

Implementation details

Architecture and initialisation The framework was implemented in PyTorch. The selected configuration uses a ResNet-18[25] encoder whose first convolution is replaced by a 7×7, stride-2 convolution accepting all 11 focal planes as input channels, followed by the four standard residual stages (feature widths 64, 128, 256, 512). The decoder is a U-Net with four upsampling blocks, each combining a transposed convolution with a skip connection from the corresponding encoder stage and two depthwise-separable convolutions, and returns a 64-channel map at half the input resolution. The segmentation head applies a 3×3 convolution, batch normalisation and ReLU followed by a 1×1 convolution producing 44 channels, reshaped to the four structures $\times$ eleven views; the view head applies global average pooling to the final encoder stage and a single linear layer producing 4×11 logits, softmax-normalised over the eleven depths independently for each structure. Per-view segmentation outputs are combined by weighting with the view distribution and summing over depths, bilinearly resampled to 800×800, and passed through PCN (equation (7)) with the containment priority RB, PB, PVS, cell body; the exclusive label map is the arg max of the resulting categorical distribution. Both encoder and decoder were initialised from the self-supervised reconstruction model described below (its reconstruction head is discarded); the comparison against ImageNet pretraining[26] and random initialisation is reported in Supplementary Table S4.

Self-supervised initialisation Initial weights were obtained by masked reconstruction of the multi-focal stacks of 17,000 unannotated oocytes, following the masked-autoencoder approach[27]. The reconstruction network shares the encoder and decoder of the segmentation model, with a reconstruction head in place of the segmentation and view heads, and its encoder was initialised from ImageNet weights before pretraining began; the in-domain arm therefore refines ImageNet features on oocyte images rather than replacing them. Masking combined two mechanisms applied to each $11 \times 800 \times 800$ stack: three of the eleven focal planes, chosen at random, were masked in their entirety, and 60% of 32×32 spatial patches were masked on each remaining plane, hiding approximately 71% of all image content. The objective was the mean squared error over masked locations only. Pretraining ran for 100 epochs with AdamW (learning rate 1.5×10^{-4}, weight decay 0.05, cosine annealing) and a batch size of 8; the checkpoint with the best reconstruction PSNR on a held-out 2% of the unannotated pool was retained. Both the pretrained encoder and the pretrained decoder were transferred to the segmentation network. A worked reconstruction example is shown in Supplementary Fig. S2.

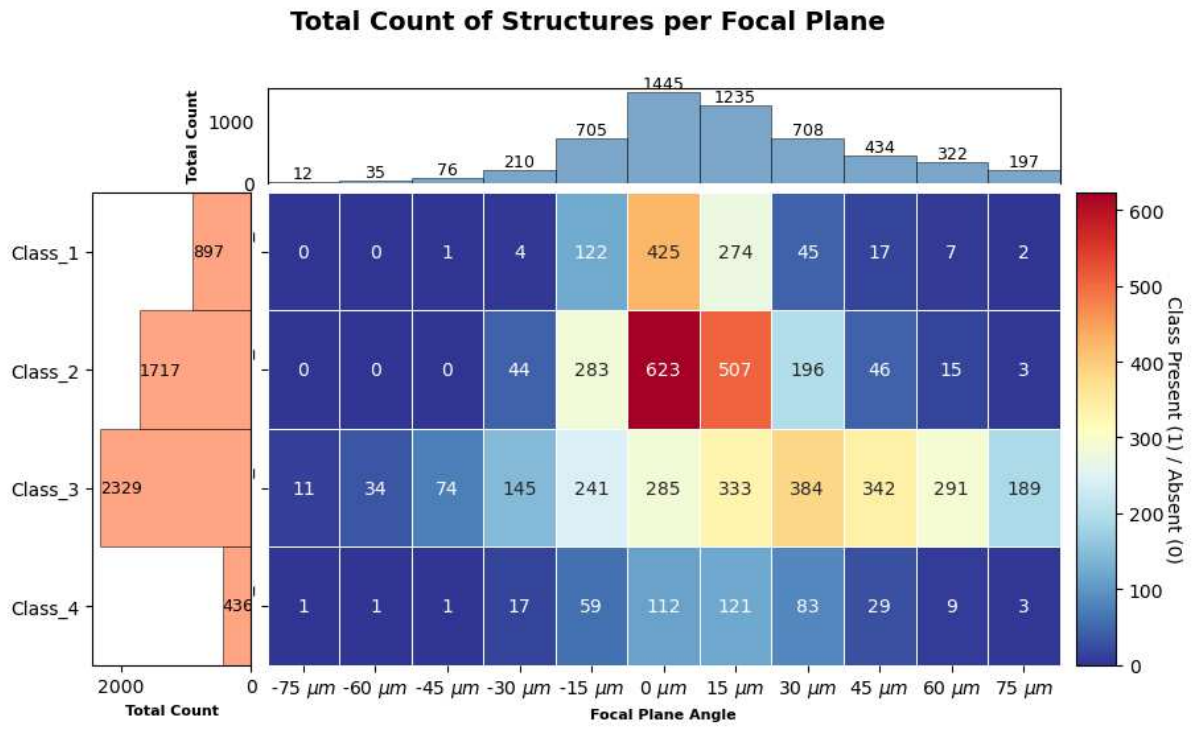

Figure 4. Distribution of annotated focal depths for each semantic structure across the full dataset. Counts are annotations per (oocyte, plane) pair, not numbers of oocytes. Classes 1 and 2 are most frequently annotated around 0–15 μm; Class 3 shows a broader distribution extending toward positive focal depths, and Class 4 a similar pattern with a narrower spread.

Training Training used the AdamW optimiser[28] with a base learning rate of 5×10^{-5}, weight decay 1×10^{-2}, and parameter-group multipliers applied to both: learning rates of 0.25, 0.40, 0.60, 0.80 and 1.00 times the base rate for the stem and the four encoder stages, respectively, 1.0 for the decoder and 2.0 for the two heads, with weight-decay multipliers of 1.0 for the stem and first three stages, 1.5 for the final stage and the decoder, and 0.5 for the heads. The learning rate was warmed up linearly from 0.1 to 1.0 of these values over the first 2 epochs and then annealed by a cosine schedule over the remaining 37 epochs to a floor of 1×10^{-7}, with a batch size of 4 and 39 epochs in total. We fixed the number of epochs and all other hyperparameters based on development-set cross-validation before retraining the final model, and all reported values are from the final epoch. No encoder parameters were frozen at any point. The random seed was fixed at 42 for all reported runs.

The total training objective was

$$\mathscr{L}(t) = \big(1 - \alpha(t)\big)\,\mathscr{L}_{\text{seg}} + \alpha(t)\,\mathscr{L}_{\text{depth}}\,, \tag{8}$$

a convex combination rather than a sum, so the weight on segmentation decreases as the depth term increases. Here $\mathscr{L}_{\text{seg}}$ combines cross-entropy and Dice[29] terms (weighted 0.05 and 0.95) computed on the PCN composite output (the cross-entropy term is a negative log-likelihood applied directly to the PCN probabilities, which are already normalised, rather than a softmax cross-entropy over logits; the Dice term is computed over the four structures only, excluding background, and is pooled over all pixels of a batch rather than averaged over oocytes), with inverse-frequency class weights recomputed on the training split of each fold (weights $\propto f_c^{-\gamma}$ with $\gamma = 1.5$, capped so that the largest does not exceed 150 times the smallest and rescaled to unit mean over the five classes; values on the full development set, for background, cell body, PVS, PB and RB respectively: 0.015, 0.099, 0.438, 2.188 and 2.261). The cap binds for the refractile body: its weight is held at the ceiling, close to that of the polar body. $\mathscr{L}_{\text{depth}}$ is the Kullback–Leibler divergence $\mathrm{KL}\big(q_{\text{target}} \,\|\, q_{\text{pred}}\big)$ between the annotation-derived and predicted depth distributions, normalised by $\ln 11$ so that it lies on a comparable scale to the segmentation term, averaged over the (oocyte, structure) pairs in which the structure is annotated; pairs with an all-zero target are excluded rather than contributing a degenerate term. The depth-loss weight $\alpha(t)$ increased linearly from 0 to 0.3 over the first 10 epochs, and was held constant thereafter. Per-depth segmentation masks are not supervised directly; segmentation supervision enters only through the aggregated composite.

Augmentation and corner-label removal Each acquisition carries a burned-in text label in the image corners. Before any augmentation, and at both training and test time, four rectangular regions of 5% of the image height by 10% of the width, one at each corner, were replaced with local background: for each region and each focal plane separately, the median and the median absolute deviation of a 12-pixel ring immediately outside the region were computed, and the region was filled with that median plus Gaussian noise scaled by 1.4826 times the deviation, so that the replacement matches the local intensity and texture rather than leaving a flat patch. Robust statistics were used so that label pixels leaking into the ring do not distort the fill. The noise was drawn with a fixed seed at test time and freely at training time. Augmentation of the training set additionally comprised horizontal and vertical flips, each with probability 0.5, synchronised between image and mask; no augmentation was applied at test time.

Metrics Dice and intersection-over-union (IoU) were computed per class by pooling predicted and annotated pixels, after taking the arg max of the PCN distribution at each pixel, rather than by averaging per-oocyte scores; a structure absent from an oocyte therefore contributes only to that structure's false-positive count and does not enter a per-oocyte average. Per-structure denominators are given in Table 1. Agreement between predicted and annotation-derived focal-depth profiles was quantified by the Jensen–Shannon distance, the square root of the Jensen–Shannon divergence. Its value ranges from 0 for identical distributions to $\sqrt{\ln 2} \approx 0.833$ for distributions with disjoint support. It was computed for each (oocyte, structure) pair in which the structure was annotated, averaged over oocytes within each structure, and then averaged over the four structures, so that each structure contributes equally regardless of how often it is annotated. Peak-depth-set overlap treats a depth as a peak of a profile if its weight is at least 0.9 times the largest weight in that profile, and counts an evaluation as a match if the predicted and annotation-derived peak sets intersect; a profile may therefore have several peaks, and a diffuse prediction has more opportunities to intersect the target set than a confident one. Two-plane overlap relaxes this criterion by checking whether either of the two highest-weighted predicted depths falls in the annotated peak set. Ranking quality was measured by average precision, the area under the precision–recall curve (PR-AUC): for each (oocyte, structure) pair, the depths at which the structure was annotated are positives, and the rest negatives, and the score measures whether the predicted distribution ranks the former above the latter. In the per-depth breakdown reported in Results, each cell is computed over the oocytes whose annotated peak set includes that depth.

Design-choice experiments Six architectural and training choices were compared during model development — the number of focal depths, how evidence across depths is combined, the source of the network's initial weights, network size, the order in which overlapping structures are resolved, and how the focal-depth target is defined — each altered one at a time against a fixed reference configuration, across the ten patient-level cross-validation splits of the development set. The reference configuration for these comparisons is ImageNet-initialised and therefore differs from the selected final configuration in its initialisation; the cross-validation performance of the selected configuration itself is reported in Supplementary Table S4. The full comparison is reported in Supplementary Table S4 and Supplementary Figs. S1 and S3.

Use of generative AI ChatGPT (GPT-6 Astra, max) was used to assist with language editing and with restructuring the manuscript text for clarity. It is not used to generate, analyse, or interpret data, or to produce figures. The authors reviewed and edited all text and take full responsibility for the content of the publication.

Results

Segmentation and focal-depth prediction on the held-out test set

The selected configuration (see Implementation details in Methods) was retrained once on the complete development set and evaluated on the independent test set of 119 oocytes, disjoint from the development set at the patient level (Methods). We fixed the architecture, hyperparameters, and number of training epochs from the ten-split development-set cross-validation before this final run; no oocyte from the test set contributed to any design comparison or to the selection of the final configuration. Test-set metrics were computed at each epoch of this final run and are plotted in Supplementary Figs. S4–S6 for completeness; the values reported below are those of the final epoch, and no checkpoint, epoch count, or architectural choice was made based on them.

Segmentation Mean Dice and mean intersection-over-union over the four annotated structures were 0.688 and 0.601; including background, the corresponding means were 0.749 and 0.678, the difference reflecting near-exact background segmentation (Dice 0.994, IoU 0.987). Accuracy differed markedly between structures (Table 1; Fig. 5): Dice was 0.969 for the cell body, 0.790 for the PVS, 0.774 for the PB, and 0.218 for the RB. For the three larger structures, sensitivity and precision were close, but they lean in opposite directions. For the PVS, sensitivity (0.847) exceeded precision (0.760): the model marks somewhat more area as PVS than the annotation does. For the PB, the order is reversed (0.764 against 0.811), so it marks somewhat less. For the RB both were low (0.191 and 0.306): the model recovered a fifth of the annotated RB area, and fewer than a third of the pixels it labelled RB had been annotated as such.

Specificity exceeded 0.99 for every structure and reached 1.000 for the RB. This says little about accuracy. Background occupies most of every image, so a structure scores near-perfect specificity simply by being predicted rarely, which is what the RB figure reflects.

The RB was annotated in 41 of the 119 test oocytes; Table 1 gives the number of oocytes behind each structure's figures.

Focal-depth prediction For each structure, the model predicts a distribution over the 11 acquired depths. Agreement with the annotation-derived distribution was quantified by the Jensen–Shannon distance (Methods), which is 0 when the two coincide and attains a maximum of $\sqrt{\ln 2} \approx 0.83$ when they share no depths. Across the test set, this distance was 0.437 (Table 1); before training, when the network's output was close to uniform across depths, it stood at approximately 0.69. Distances were smaller for the PVS and cell body (0.32 and 0.35) than for the PB and RB (0.53 and 0.57), an ordering that was established within the first ten epochs and stable thereafter (Supplementary Fig. S6).

A simpler question is whether the model picks out the right depth. For each oocyte and structure, we took the depth or depths carrying the highest predicted weight and asked whether any of them was also among the most heavily annotated depths (Methods). By this criterion the model was right in 53.5% of cases overall: 78% for the cell body, 62% for the PVS, 41% for the PB and 32% for the RB (Table 1; broken down by depth in Fig. 6a).

This ranks the cell body ahead of the PVS, whereas the Jensen–Shannon distance ranks them the other way round. The two measures ask different things. The Jensen–Shannon distance scores the shape of the whole profile; this criterion scores only its peak. The cell body is annotated at a single depth in 99% of oocytes (Supplementary Table S7), so its target has one sharp peak: easy to hit, but hard to match in shape when the prediction spreads its weight over neighbouring depths. The PVS is the reverse.

Peak overlap has two weaknesses, and we treat it as secondary. It depends on a threshold, and it favours diffuse predictions: a profile with several near-equal maxima has more chances to overlap the annotated set than one with a single clear peak.

Allowing the model two guesses instead of one changes the picture considerably. If either of the two highest-weighted predicted depths counts as a hit, the model is right in 77% of cases, against 53.5% with one (Table 1; Supplementary Fig. S6). Naming one depth per structure is therefore right about half the time; naming two is right about three times in four.

A ranking measure asks a looser question: does the model rank the depths an embryologist annotated above those they did not? This is the area under the precision–recall curve (PR-AUC; Methods). It needs no threshold and does not require the model to single out one depth. A random ranking of the eleven depths would score about 0.27 when one depth is annotated and about 0.42 when three are (Methods). The model scored 0.79 overall, and 0.90, 0.92, 0.68 and 0.60 for the cell body, PVS, PB and RB respectively (Table 1; Fig. 6b). Every structure is well above chance, including the RB: it is usually annotated at one or two depths, for which chance lies around 0.3, so its 0.60 is roughly double.

The polar body shows the pattern most clearly. Its top depth is right in only 41% of oocytes, yet its PR-AUC is 0.68, and it stays between 0.52 and 0.81 wherever in the stack the polar body lies (Fig. 6b). One reason is that a polar body is usually annotated across two or three depths of similar weight (Supplementary Table S7), so picking out the single best one is hard even when the model has found the right region. In short, the model is better at pointing to where a structure is in the stack than at naming the one best depth.

To gauge overfitting, we compared the final model's performance on its own training data with its performance on the held-out set (Supplementary Figs. S4 and S6). For segmentation the gap was modest: Dice 0.81 against 0.75 and mean IoU 0.74 against 0.68, both averaged over all five classes. For the Jensen–Shannon distance there was almost no gap (0.42 against 0.44). Peak overlap showed the largest gap of any measure, 70% against 53.5%, consistent with its being the most sensitive to small shifts in where the model places its highest weight.

Held-out performance also matched what cross-validation predicted. Across the ten development-set splits, the selected configuration reached a mean IoU of 0.657 (Supplementary Table S4), against 0.678 on the test set, again averaging all five classes.

Structure-specific performance and annotation overlap

The four structures are not separate regions of the image: in the two-dimensional composite projection, the RB lies within the cell body, and the PB lies against the oocyte membrane within the PVS, so a single image location can belong to more than one structure at once. Across the 778 oocytes of the development set, 93% of the area annotated as PB and 73% of that annotated as RB is shared with at least one other structure, compared with under 13% for the cell body and PVS (Supplementary Tables S1 and S2). The pattern of overlap matches known oocyte morphology — the RB overlaps almost exclusively with the cell body, and the PB is divided between the cell body and the PVS — which supports the internal coherence of the annotations themselves.

This overlap, combined with size and rarity, is a plausible explanation for the performance gap between structures, although the three factors coincide in this dataset rather than varying independently and were not isolated. The cell body occupies 82.5% of the unique annotated image area, is only lightly contested (4.85%), and is segmented almost exactly. The PB is much smaller and heavily entangled with its neighbours, yet it is segmented well, which indicates that substantial overlap is not by itself an

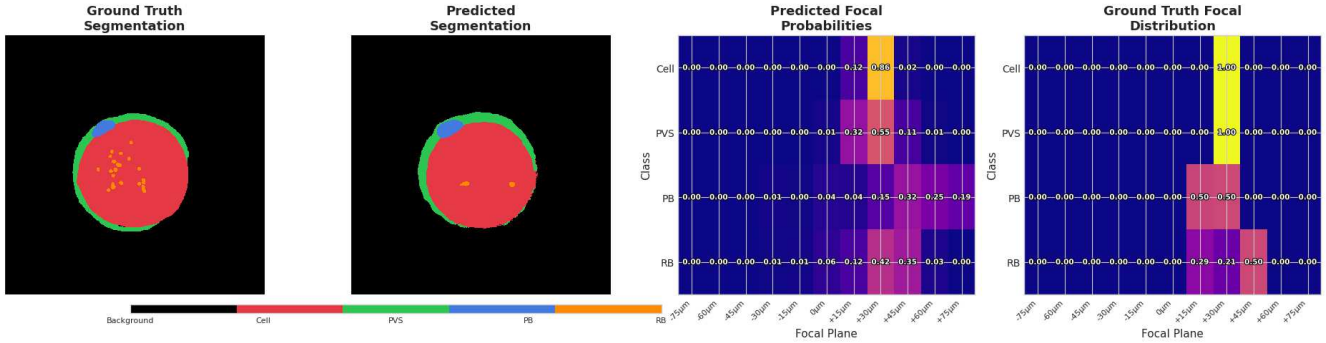

Figure 5. Both outputs of the framework for one test oocyte. **a**, Annotated segmentation. **b**, Predicted segmentation. **c**, Predicted distribution across focal depths for each structure. **d**, Annotation-derived distribution. The cell body, perivitelline space and polar body are recovered closely. Most of the annotated refractile body, a scattering of small inclusions, is not predicted. In this oocyte, the annotation places the cell body and PVS at a single depth (+30 μm), the PB across two (+15 and +30 μm) and the RB across three (+15 to +45 μm). The predicted distributions put their weight at those depths but spread it over neighbouring planes, most visibly for the PB. This is a qualitative example, not an aggregate result.

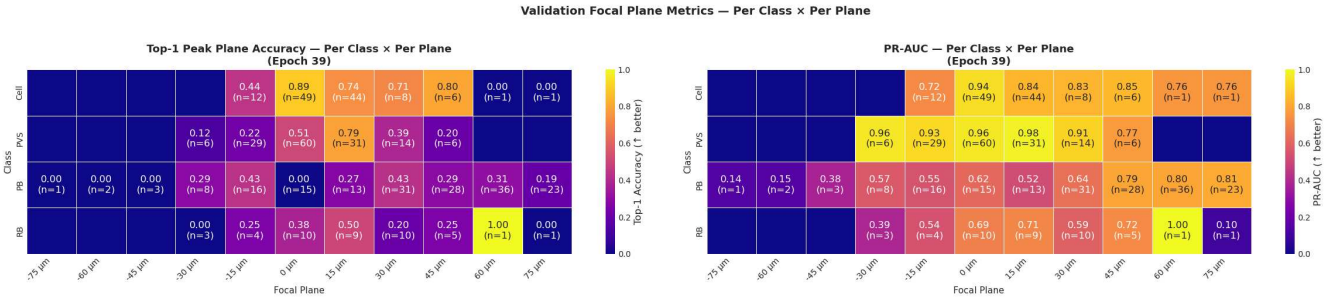

Figure 6. Focal-depth prediction by structure and depth, on the held-out test set. Each column groups the oocytes by the depth at which a structure's annotation peaks, and the number of oocytes in each group is printed in the cell. **a**, How often the model's predicted peak also falls at that depth. **b**, For the same oocytes, precision–recall AUC across all eleven depths, which measures how well the model ranks the depths at which the structure was annotated above those at which it was not; it depends on no threshold. Blank cells are depths at which the structure's annotation never peaks; they are not failures. Cells resting on very few oocytes are not performance estimates; the refractile body at +60 μm rests on one oocyte. Training curves for both measures are in Supplementary Fig. S6.

Table 1. Performance of the selected configuration on the independent test set (119 oocytes). *n* is the number of test oocytes in which each structure was annotated. **Segmentation** (top): measures are pooled over all pixels of the test set. Specificity is shown for completeness but says little here, because background fills most of every image. **Focal-depth prediction** (bottom): measures are averaged over the oocytes in which each structure is annotated, then over the four structures, so the overall row has no single *n*; background is not a focal-depth target. *Peak overlap* is the proportion of cases in which the model's most heavily weighted depths (within 90% of its maximum) include one of the most heavily annotated depths; allowing the model's two highest-weighted depths raises it to 77% overall. *PR-AUC* measures how well the model ranks annotated depths above unannotated ones; a random ranking scores about 0.27 to 0.42, depending on how many depths are annotated. Development-set cross-validation results and the full design-choice comparison are in Supplementary Table S4.

Segmentation	*n*	**Dice**	**IoU**	**Sens.**	**Prec.**	**Spec.**
Cell body	119	0.969	0.941	0.975	0.964	0.995
Perivitelline space	115	0.790	0.660	0.847	0.760	0.992
Polar body	117	0.774	0.647	0.764	0.811	0.999
Refractile body	41	0.218	0.155	0.191	0.306	1.000
Mean, four structures		0.688	0.601			
Background		0.994	0.987			
Mean, five classes		0.749	0.678			

Focal-depth prediction	*n*	**JS distance**	**Peak overlap**	**PR-AUC**
Cell body	119	0.35	78%	0.90
Perivitelline space	115	0.32	62%	0.92
Polar body	117	0.53	41%	0.68
Refractile body	41	0.57	32%	0.60
Overall	—	0.437	53.5%	0.79

obstacle when enough annotated examples are available. The RB differs in all three respects at once: annotated in only 266 of the 897 oocytes of the full cohort, occupying a quarter of one percent of annotated area, and with three quarters of that area entangled with the cell body, so its visual evidence is both rare and ambiguous.

The model responds by predicting the RB conservatively, recovering a fifth of its annotated area at a precision of 0.306 (Fig. 5). The model's focal-depth output for the RB is better than its segmentation. Although it outlines only a fifth of the RB's area, it still ranks the depths at which the RB was annotated above the others, with a PR-AUC of 0.60, about twice what random ranking would give (Fig. 6). In other words, the model often indicates where in the stack to look for an RB even when it cannot delineate it. This is an average over the 41 test oocytes with an RB; we did not check whether it holds in the individual oocytes where segmentation fails.

This pattern is clinically consequential rather than incidental: the PB and RB are, of the four structures, those most consistently linked in the literature to blastocyst formation and implantation outcome[4,5]. The framework performs well on one of these two structures and poorly on the other. Several factors may contribute to the RB result: limited annotation volume; anatomical overlap; the composite target construction; and annotation consistency for so small a structure. None of these can be excluded on the present evidence.

Discussion

This study asked whether one model can segment four oocyte structures from a multi-focal stack while also predicting, for each structure, the depths at which an embryologist annotated it as clearest. On patient-disjoint held-out data, segmentation was accurate for the cell body and good for the PVS and PB, but poor for the RB. The predicted focal-depth profiles agreed with the annotation more closely for the cell body and PVS than for the PB and RB. Two design choices mattered: supplying the full focal stack rather than one plane, and initialising the network by self-supervised learning on unannotated oocyte images.

Published accuracy figures for single-plane oocyte segmentation are higher than ours, but they are not comparable. Each was obtained against an annotation drawn on the same plane the model received, on different target structures and a different cohort. Our target is a composite across eleven depths, and two of its four structures occupy less than 4% of the annotated area between them, so the mean across structures is harder to attain and is pulled down further by the refractile body. We did not train any published architecture on this cohort, so we cannot claim superiority over one. Instead, we report two internal variants that isolate what single-plane and generic multi-focal methods each lack, with architecture, training and evaluation held fixed.

The single-central-plane variant sees one depth, which approximates a conventional single-plane method, though with a single view its focal-depth loss is identically zero, so it is a segmentation-only model; it performed worse ($\Delta - 0.0373$ mean IoU). The equal-weighting variant sees the full stack but does not learn which depths matter for which structure, which approximates a generic multi-focal model; it was only marginally worse ($\Delta - 0.0035$ mean IoU), a difference well within the variation between cross-validation splits (s.d. 0.0116 for the reference; Supplementary Table S4).

The equal-weighting comparison bears directly on what we claim. Learning how much weight to give each depth did not make the segmentation more accurate than simply averaging all depths equally. What learning the weights adds is a second output: for each structure in each oocyte, a profile of the depths at which it was most likely annotated. Equal averaging produces no such profile, and the profile comes at no measurable cost in segmentation accuracy. Our claim is therefore not that learned weighting improves segmentation, but that it provides focal-depth information alongside a segmentation of comparable accuracy.

Two findings are worth testing in other datasets. The first is that giving the model all eleven depths, rather than only the central one, helped the smallest structures most: refractile-body Dice rose from 0.12 to 0.18, while cell-body Dice rose only from 0.93 to 0.95 (Supplementary Table S4). This supports the premise the framework rests on, that the depths a single-plane image leaves out contain useful information, especially for small structures that are sharp at only a few of them. It is the clearest evidence we have for that premise, though it comes from one architecture and one cohort.

The second is that how the network's weights were initialised mattered more than any other choice we examined. The best initialisation came from pretraining the network, without any annotations, to fill in deliberately hidden parts of unannotated focal stacks (masked reconstruction). Transferring both halves of this pretrained network raised development-set mean IoU by 0.1074 over random initialisation, with all other training settings matched. It also beat standard ImageNet initialisation, by 0.0553, but two caveats apply to that comparison. The two arms used different learning-rate settings, and the pretrained network had itself started from ImageNet weights before it saw any oocyte images. The oocyte pretraining therefore built on ImageNet features rather than replacing them, and we cannot say what it would achieve from a random start.

What the results do suggest is that 17,000 unannotated oocyte acquisitions add substantially to what a general-purpose image collection provides on its own. Such acquisitions accumulate as a by-product of routine practice, so this is a cheap approach to test elsewhere. It also outperformed the other route to higher accuracy we tried, a larger network (ResNet-50, $+0.0196$ mean IoU). We did not try combining the two.

The polar body is the clearest test of whether the focal-depth output reflects the individual oocyte. Across the cohort it can appear at any depth: it is annotated at every depth from -75 to $+75\,\mu$m, and no single depth carries more than 0.16 of the average target in the development set (Supplementary Table S8). In any one oocyte, however, it is annotated at only two to three depths on average (Supplementary Table S7). No fixed set of depths would therefore work for every oocyte; to do well, a model has to judge each one. The model ranks the polar body's annotated depths well wherever they lie: PR-AUC is 0.68 overall, and between 0.52 and 0.81 when oocytes are grouped by the depth of their annotated peak (Fig. 6). This suggests that the predicted profile responds to the individual oocyte rather than repeating a population average, although we did not compare the two directly.

The cell body and PVS are the opposite case. Both are annotated at 0 to $+15\,\mu$m in most oocytes, so a model that ignored the individual oocyte and always returned the population profile would already score well on them. Their agreement figures should be read with that in mind.

The focal-depth measures ask progressively looser questions, and the model does better as the question loosens. The strictest asks whether the model's top depths match an annotated peak: this holds about half the time (53.5%), and it is the measure that generalises worst, falling from 70% on the training set. Allowing the model two guesses instead of one raises this to 77%. Looser still, PR-AUC asks only whether annotated depths are ranked above unannotated ones, and gives credit when they come second or third: it is 0.79 overall and 0.60 even for the refractile body, well above chance in both cases. The Jensen–Shannon distance, which compares the whole profile, barely changes between training and held-out data (0.42 against 0.44). The practical reading is that the output is more dependable at indicating a region of the stack worth examining than at naming one best depth. That is the weaker claim, and it is the one an interface should be built on.

Refractile-body segmentation remained far weaker than that of the other three structures, and is the principal limitation of this work. Three properties may contribute: it is annotated in only 266 of 897 oocytes, occupies a quarter of one percent of annotated area, and shares three quarters of that area with the cell body. We did not isolate their effects, and cannot: in this dataset the three vary together. Nor can we exclude limitations of the model, of the composite target construction, or of annotation consistency for a structure this small. Whether more RB annotation would close the gap is a hypothesis, not a result, though it could be tested directly by measuring performance against annotation volume.

The RB result is not wholly negative, however. Although the model outlines only a fifth of the RB's annotated area, its focal-depth output for the RB remains informative: it ranks the depths at which the RB was annotated above the others with a PR-AUC of 0.60, about twice what random ranking would give. The model therefore often indicates where in the stack to look

for an RB even when it cannot outline it. This is an average over the 41 test oocytes with an RB; we did not check whether it holds in the individual oocytes where segmentation fails. More broadly, pixel-level Dice is a demanding criterion for so small a structure, and oocyte-level detection — whether an RB is present at all — may be the more clinically meaningful endpoint. We did not evaluate it.

The framework produces two outputs that could support manual oocyte assessment: a composite segmentation, and, for each structure, a distribution over focal depths. One possible interface would show the predicted masks alongside these depth weights, so that an embryologist could decide which depths to open first. It could be organised by structure, as a shortlist of depths worth reviewing for each, or by depth, as an overlay showing which structures are predicted at each one (Fig. 7). On our results, a shortlist of two depths per structure would include an annotated peak in about three cases in four. Either way, the interface would support the multi-depth inspection embryologists already perform rather than replace it.

We have not tested either output in practice, so several cautions apply. First, any interface should keep all eleven depths available to the embryologist. The predicted weights rank the depths by how likely each is to show a structure clearly, but their values are not probabilities: a weight of 0.4 does not mean a 40% chance, and we did not test how well the weights match observed frequencies. Second, the interface should flag where the model is less reliable: its depth predictions for the polar body, and its segmentation of the refractile body. Third, it remains unknown whether such an interface would make review faster, make embryologists agree more often, or change any clinical decision. Finally, the framework does not assess oocyte quality. We did not relate its outputs to oocyte grading, fertilisation or blastocyst development, and on present evidence its refractile-body segmentation is not accurate enough for clinical use.

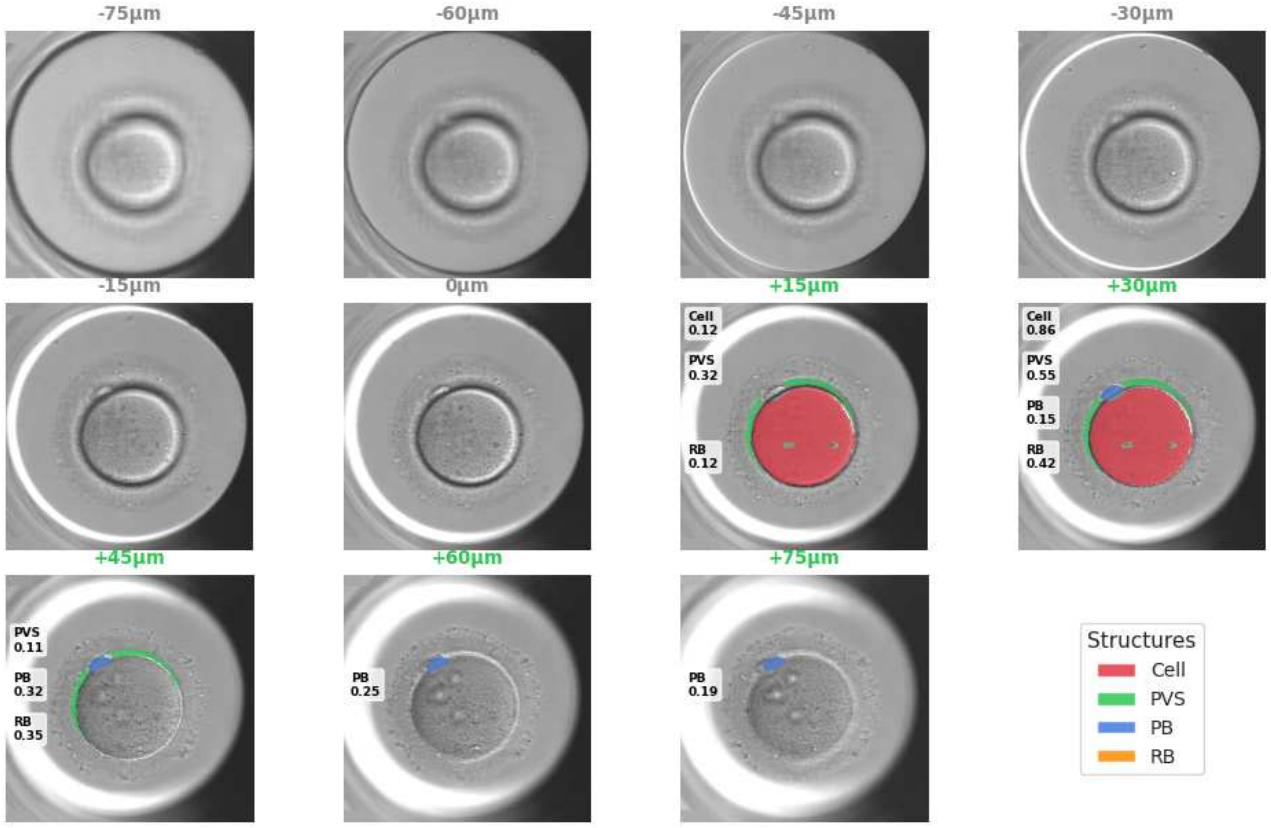

Figure 7. Predicted structures overlaid on the eleven focal depths of one test oocyte. Depths at which the model predicts at least one structure are labelled in green, and the predicted weight for each structure is shown beside the image. For this oocyte the predictions fall between +15 and +75 μm, and each structure is predicted to be most visible at a different depth. The figure illustrates the per-depth view described in the text; it is a qualitative example from one oocyte, not an aggregate result.

These results establish the approach on a carefully annotated single-centre cohort, and they point to clear next steps. Validation in other laboratories and imaging set-ups is the natural first extension, together with a larger test set that would allow confidence intervals, particularly for the refractile body, which was present in 41 of the 119 test oocytes. Measuring agreement between embryologists on the same images would show how the model's accuracy compares with clinical consistency, which

is the benchmark that matters most for adoption. Two further analyses would sharpen the focal-depth findings: annotating every depth at which a structure is visible, rather than only the clearest, and comparing the predicted profiles with a population average for each structure. The latter matters most for the cell body and PVS, which sit at similar depths in most oocytes, and would confirm how much of the agreement reflects the individual oocyte; the polar body, whose depth varies widely, already offers a more stringent test. Finally, the design choices were tested individually, and combining the most effective of them, notably self-supervised initialisation with a larger network, may yield further gains. Focal-depth prediction is hardest for the polar and refractile bodies, the structures whose position varies most between oocytes, and these are where improvements would bring the greatest practical benefit.

Conclusion

In this single-centre study, a mixture-of-views network produced two outputs from an 11-depth focal stack: a composite segmentation of four oocyte structures, and, for each structure, a profile of the depths at which it is most clearly seen. The depth profile comes at no measurable cost in segmentation accuracy. Segmentation was accurate for the cell body, perivitelline space and polar body, and limited for the refractile body. Supplying the full stack rather than a single central depth improved segmentation most for the smallest structures, and self-supervised initialisation on unannotated oocyte images was the most beneficial of the design choices examined.

In practice, the depth profile tells an embryologist where in the stack to look. For each structure, the model's two highest-ranked depths include one the embryologist annotated in about three of every four oocytes. An embryologist could therefore start with these two depths instead of scanning all eleven, while still having the full stack available. Its value is greatest where guidance is hardest to provide in advance: the polar body's clearest depths vary widely from oocyte to oocyte, yet the model ranks them well wherever they lie. Even for the refractile body, whose outline the model captures poorly, the profile still points to the right depths far more often than chance, and more annotation is a clear route to test for better segmentation. Validation on other cohorts, comparison with agreement between embryologists, and evaluation within clinical workflow are the next steps towards practical use.

Data availability

The data that support the findings of this study are available from UZ Brussels but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are, however, available from the authors upon reasonable request and with permission of UZ Brussels.

Code availability

The code used to train and evaluate the models is available to editors and reviewers on request and will be released publicly on GitHub, upon publication.

References

1. Pandit, S. & Sharma, R. Non-invasive assessment of human oocytes and embryos in assisted reproduction: review on present practices and future trends. *Med. J. Armed Forces India* **78**, 7–16 (2022).
2. Rienzi, L., Vajta, G. & Ubaldi, F. Predictive value of oocyte morphology in human IVF: a systematic review of the literature. *Hum. Reprod. Update* **17**, 34–45 (2011).
3. Firuzinia, S., Afzali, S. M., Ghasemian, F. & Mirroshandel, S. A. A robust deep learning-based multiclass segmentation method for analyzing human metaphase II oocyte images. *Comput. Methods Programs Biomed.* **201**, 105946 (2021).
4. Takahashi, H. *et al.* Clinical outcomes of MII oocytes with refractile bodies in patients undergoing ICSI and single frozen embryo transfer. *Reprod. Med. Biol.* **19**, 75–81 (2020).
5. Bartolacci, A. *et al.* Does morphological assessment predict oocyte developmental competence? A systematic review and proposed score. *J. Assist. Reprod. Genet.* **39**, 3–17 (2022).
6. Targosz, A., Przystałka, P., Wiaderkiewicz, R. & Mrugacz, G. Semantic segmentation of human oocyte images using deep neural networks. *Biomed. Eng. Online* **20**, 40 (2021).
7. Fjeldstad, J. *et al.* An artificial intelligence tool predicts blastocyst development from static images of fresh mature oocytes. *Reprod. Biomed. Online* **48**, 103842 (2024).
8. Fjeldstad, J. *et al.* Segmentation of mature human oocytes provides interpretable and improved blastocyst outcome predictions by a machine learning model. *Sci. Rep.* **14**, 10569 (2024).

9. Ganguly, A., Jain, S. & Watchareeruetai, U. Amortized variational inference: a systematic review. Preprint at https://doi.org/10.48550/arXiv.2209.10888 (2022).

10. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. Preprint at https://doi.org/10.48550/arXiv.1505.04597 (2015).

11. Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. Rethinking atrous convolution for semantic image segmentation. Preprint at https://doi.org/10.48550/arXiv.1706.05587 (2017).

12. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. Preprint at https://doi.org/10.48550/arXiv.1802.02611 (2018).

13. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).

14. Conze, P.-H., Andrade-Miranda, G., Singh, V. K., Jaouen, V. & Visvikis, D. Current and emerging trends in medical image segmentation with deep learning. *IEEE Trans. Radiat. Plasma Med. Sci.* **7**, 545–569 (2023).

15. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).

16. Cohen, E. & Uhlmann, V. Aura-net: robust segmentation of phase-contrast microscopy images with few annotations. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* 640–644 (IEEE, 2021).

17. Targosz, A., Myszor, D. & Mrugacz, G. Human oocytes image classification method based on deep neural networks. *Biomed. Eng. Online* **22**, 92 (2023).

18. Daood, A., Ribeiro, E. & Bush, M. Sequential recognition of pollen grain Z-stacks by combining CNN and RNN. In *Proceedings of the International Florida Artificial Intelligence Research Society Conference (FLAIRS)* (AAAI Press, 2018).

19. Manescu, P. *et al.* Deep learning enhanced extended depth-of-field for thick blood-film malaria high-throughput microscopy. Preprint at https://doi.org/10.48550/arXiv.1906.07496 (2019).

20. Souda, A. B. *et al.* Enhancing embryo stage classification with multi-focal plane imaging and deep learning. Preprint at https://doi.org/10.1101/2025.03.21.644547 (2025).

21. Jonaitis, D., Raudonis, V., Drejeriene, E., Kozlovskaja-Gumbriene, A. & Salumets, A. Soft optical sensor for embryo quality evaluation based on multi-focal image fusion and RAG-enhanced vision transformers. *Sensors* **26**, 1441 (2026).

22. Pelapur, R. *et al.* Multi-focus image fusion using epifluorescence microscopy for robust vascular segmentation. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 4735–4738 (IEEE, 2014).

23. Pei, R. *et al.* TransFusion-net for multifocus microscopic biomedical image fusion. *Comput. Methods Programs Biomed.* **240**, 107688 (2023).

24. Xu, C., Tao, Y., Guo, X. & He, H. High-fidelity and efficient particle microscopy via regional focus search and adaptive focus stacking. *Sci. Rep.* **16**, 5755 (2026).

25. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016).

26. Deng, J. *et al.* ImageNet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).

27. He, K. *et al.* Masked autoencoders are scalable vision learners. Preprint at https://arxiv.org/abs/2111.06377 (2021).

28. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. Preprint at https://arxiv.org/abs/1711.05101 (2019).

29. Milletari, F., Navab, N. & Ahmadi, S.-A. V-Net: fully convolutional neural networks for volumetric medical image segmentation. Preprint at https://doi.org/10.48550/arXiv.1606.04797 (2016).

Acknowledgements

This work was funded by the Brussels-Capital Region - Innoviris through the FROSCT project (grant 2024-JRDIC-4b).

Author contributions

G.K.K. conceived and implemented the methodology, conducted the experiments, analysed the results, and drafted the manuscript. S.W. contributed to the investigation and manuscript review. K.W. and M.D.V. contributed clinical expertise and data annotation. A.D. and T.L. supervised the work and contributed to the methodology and manuscript review. All authors reviewed and approved the manuscript.

Additional information

Supplementary information accompanies this paper.

Competing interests: The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ScientificReportSupplementary.pdf