Supplementary Information for: 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], 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

This document supports the main text with five notes:

- Supplementary Note S1 derives the bound on the error introduced by using eleven discrete focal depths, referred to in Methods.
- Supplementary Note S2 gives the full breakdown of how much the four structures overlap one another.
- Supplementary Note S3 reports the complete results of the six design-choice experiments that the main text summarises.
- Supplementary Note S4 gives the composition of the data splits, the image-normalisation constants, statistics on how many depths each structure is annotated at, and the chance level of the PR-AUC measure.
- Supplementary Note S5 shows the training curves of the selected model.

S1 Bound on the discretisation error

The continuous mixture-of-views (MoV) expectation used in Methods is replaced by a finite mixture over the acquired focal planes. Write

$$t(z) \triangleq \mathbb{P}(T_s(u) = 1 \mid Z_s = z, \mathbf{X}), \qquad I = \int_{-F}^{F} t(z)\, p(z \mid \mathbf{X})\, \mathrm{d}z,$$

and partition $[-F, F]$ into disjoint measurable intervals $\{B_k\}_{k=1}^{K}$ of length $\Delta_k = \|B_k\|$ whose union is $[-F, F]$. Let

$$w_k = \int_{B_k} p(z \mid \mathbf{X})\, \mathrm{d}z, \qquad \text{so that} \qquad \sum_{k=1}^{K} w_k = 1,$$

be the probability mass of interval B_k — the same quantity the main text denotes $\mathbb{P}(z_s = k \mid \mathbf{X})$ — and let $z_k \in B_k$ be a representative depth. The finite mixture actually used is

$$\tilde{I} = \sum_{k=1}^{K} w_k \, t(z_k). \tag{S1}$$

Assume only that t is Lipschitz continuous on $[-F, F]$ with constant L_t; note that $t(z) \in [0, 1]$ by construction. Then

$$\begin{aligned}
|I - \tilde{I}| &= \left| \sum_{k=1}^{K} \int_{B_k} \big(t(z) - t(z_k)\big) \, p(z \mid \mathbf{X}) \, \mathrm{d}z \right| \\
&\leq \sum_{k=1}^{K} \int_{B_k} \big|t(z) - t(z_k)\big| \, p(z \mid \mathbf{X}) \, \mathrm{d}z \\
&\leq L_t \sum_{k=1}^{K} \int_{B_k} |z - z_k| \, p(z \mid \mathbf{X}) \, \mathrm{d}z \\
&\leq L_t \Delta \sum_{k=1}^{K} w_k = L_t \Delta,
\end{aligned}$$

where $\Delta = \max_k \Delta_k$ and the last inequality uses $|z - z_k| \leq \Delta_k \leq \Delta$ for every $z \in B_k$. Hence

$$|I - \tilde{I}| \;\leq\; L_t \Delta, \tag{S2}$$

and the error vanishes as the partition is refined ($\Delta \to 0$).

Two remarks. First, no Lipschitz condition on the density p is required, because the mixture weights are exact interval masses rather than point-density approximations $p(z_k)\Delta_k$; the bound is correspondingly tighter than the Riemann-sum bound $L_t\Delta + 2FL_p\Delta$ that the alternative construction yields, and it matches the quantity the model actually computes.

Second, this bound concerns only the discretisation of a fixed continuous model. It does not bound the estimation error of the learned network, does not establish that the predicted distributions are calibrated, and does not establish that the 11 acquired planes are sufficient for any particular structure. It states that the finite-view formulation is a controlled approximation of the continuous one, not that the approximation is adequate for a given clinical purpose.

S2 Structural overlap between oocyte structures

The four annotated structures are not separate regions of the image: the refractile body (RB) lies within the cell body, and the polar body (PB) lies against the oocyte membrane within the perivitelline space (PVS), so a single image location can belong to more than one structure at once. The main text (Results, "Structure-specific performance and annotation overlap") reports the headline overlap percentages and their implication for interpreting per-structure performance; this section gives the complete pairwise breakdown.

For each oocyte in the 778-oocyte development set, the image locations at which each structure had been annotated on at least one of the 11 focal depths were recorded. A location was termed *contested* if more than one structure had been annotated there. Of the 81,177,577 annotated image locations in the development set, 3,959,921 (4.88%) are contested overall; this aggregate figure is of

Supplementary Table S1: Proportion of each structure's annotated area that is shared with at least one other structure, across the 11 focal depths. Percentages of total annotated area are given for reference: these are the size and rarity figures quoted in the main text. They sum to more than 100% because a contested location is counted under every structure annotated there.

Structure	Annotated area (px)	% of total	Contested (px)	% contested
Cell body	67,002,045	82.54%	3,246,936	4.85%
Perivitelline space	15,117,639	18.62%	1,909,218	12.63%
Polar body	2,875,100	3.54%	2,675,736	93.07%
Refractile body	204,794	0.25%	150,032	73.26%

little use on its own, since the cell body alone accounts for 82.5% of annotated area and is scarcely affected. Considered structure by structure, the position is quite different (Supplementary Table S1).

Examined pair by pair (Supplementary Table S2), the pattern of overlap corresponds to known oocyte morphology: the RB overlaps almost exclusively with the cell body and essentially never with the PVS, as expected of a structure lying within the ooplasm rather than outside it; overlap of the PB is divided between the cell body and the PVS, consistent with its position against the oocyte membrane. This consistency with known anatomy supports the internal coherence of the annotations.

Supplementary Table S2: Number of image locations annotated as belonging to both structures in a pair, across the full development set.

Pair of structures		Shared locations
Cell body	Polar body	1,962,751
Cell body	Perivitelline space	1,197,047
Perivitelline space	Polar body	769,938
Cell body	Refractile body	149,218
Polar body	Refractile body	5,118
Perivitelline space	Refractile body	9

The RB is the structure with the least unshared image evidence: three quarters of its area is also cell body. It is also the structure that gained most from two design choices in Supplementary Note S3: from supplying the full focal stack rather than one depth, and from self-supervised initialisation. The two observations may be connected, but we cannot show that they are. The RB is also the smallest and the least often annotated structure, and in this dataset overlap, size and annotation volume vary together, so their effects cannot be separated.

S3 Design-choice experiments: full results

The experiments reported here follow the machine-learning convention of the term *ablation*: each alters or removes a single component of the model, holding everything else fixed, so that the contribution of that component can be isolated. The usage denotes a controlled comparison and has no connection to tissue ablation.

S3.1 Experimental design

Six architectural and training choices of the proposed framework were investigated before fixing the final configuration used in the main text: 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 experiment varied one choice at a time while keeping all other settings unchanged, with one exception: the comparison against ImageNet initialisation also differed in learning-rate settings (Supplementary Note S3.3). Performance was assessed across ten patient-level cross-validation splits of the development set, with every variant trained and evaluated on the same splits; no oocyte of the held-out test set entered any of these comparisons.

The reference configuration for every comparison uses an ImageNet-pretrained ResNet-18 encoder adapted to process the full $11 \times 800 \times 800$ focal stack, together with all 11 focal depths, learned weighting of those depths, containment-ordered overlap resolution, and an area-weighted focal-depth target. It therefore differs from the selected final configuration, which is initialised from self-supervised weights learned on unannotated oocyte images; the selected configuration's own cross-validation performance is given in the first block of Supplementary Table S4, and is the figure the main text compares with test-set performance. The loss configuration, class weighting, and augmentation used for every variant are as described in the main text's "Implementation details" subsection, and are identical across arms except for the specific choice under comparison. The different choices, the questions they address, and the corresponding variants are summarised in Supplementary Table S3.

Metric Segmentation accuracy was measured by mean intersection-over-union (IoU) and by the Dice coefficient of each structure, both ranging from 0 to 1. Throughout this section, *mean IoU* averages over all five classes, background included; on the test set this quantity is 0.678 (main text, Table 1). The main text also reports a mean over the four structures only, 0.601 on the test set. That this second figure almost equals the reference configuration's 0.6017 below is a coincidence: the two are different quantities, on different data. Per-structure Dice is reported alongside the mean because the cell body alone contributes more than 80% of annotated pixels (Supplementary Table S1) and therefore dominates any class mean.

Statistical analysis Differences from the reference configuration were tested with two-sided paired Wilcoxon signed-rank tests across the ten matched splits ($n = 10$), chosen because fold-level differences cannot be assumed normal at this sample size. Significance is reported in this work as $p < 0.01$, $p < 0.05$ or not significant. The training sets of the ten splits overlap and thirteen configurations were compared against a common reference; these comparisons are therefore reported as exploratory, and the significance levels should not be read as independent confirmatory tests. No comparison is interpreted as evidence of equivalence on the grounds of non-significance.

S3.2 How many focal depths are needed

Reducing the number of focal depths consistently decreased segmentation accuracy, from a mean IoU of 0.6017 ± 0.0116 with all 11 depths to 0.5643 ± 0.0105 with a single central plane — the closest equivalent to conventional single-shot imaging. Each reduced-depth configuration performed worse than the full stack (Supplementary Fig. S1; $p < 0.01$ at 1, 3 and 5 depths; $p < 0.05$ at 7). The benefit of multi-focal imaging fell mainly on the smallest structure. Going from the central depth

Supplementary Table S3: The six components examined, the question each addresses, and the variants compared.

Component	Question addressed	Variants compared
Number of focal depths	Does the full stack beat fewer depths?	1, 3, 5, 7, 11 (reference)
Combining depths	Does learning how to weight the depths beat fixed alternatives?	learned (reference) vs. equal averaging vs. single depth
Initial weights	How much does the starting point matter?	random vs. ImageNet (reference) vs. oocyte images, encoder vs. oocyte images, both halves
Network size	Does added capacity help at this sample size?	ResNet18 (reference) vs. ResNet34 vs. ResNet50
Overlap resolution	Does anatomical containment improve accuracy for the polar and refractile bodies?	containment (reference) vs. reversed order
Focal-depth target	Does weighting by visible area beat presence alone?	area-weighted (reference) vs. presence

alone to all eleven depths raised RB Dice from 0.12 to 0.18, but cell-body Dice only from 0.93 to 0.95.

In the reduced-depth variants the model is constructed with the corresponding number of view slots, so the single-plane variant has a degenerate view distribution (one plane of weight one) and its focal-depth loss term is identically zero. That arm is therefore a segmentation-only model rather than the full framework restricted to one plane, which should be borne in mind when reading it as a proxy for conventional single-plane methods.

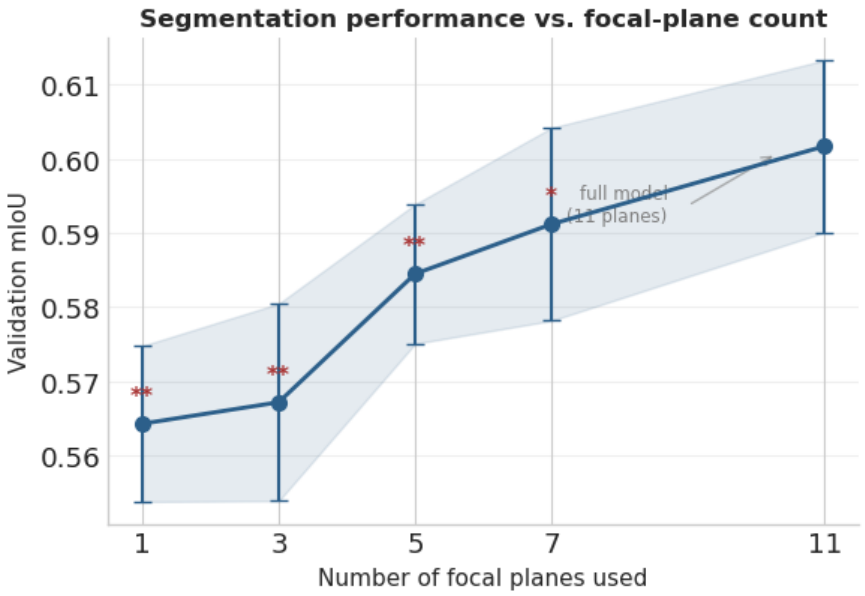
S3.3 Where the model's initial weights come from

Model initialisation was specifically investigated by comparing three sources of initial weights: random initialisation, ImageNet pretraining, and self-supervised pretraining on unannotated oocyte focal stacks. For the latter, the network was trained without annotations to reconstruct masked regions of the focal stack (Supplementary Fig. S2). Because this procedure trains both encoder and decoder, transfer of either the pretrained encoder alone or both pretrained components was evaluated.

The self-supervised model was itself initialised from ImageNet weights before reconstruction training began. The "oocyte images" arm is therefore ImageNet features refined on in-domain data, not an alternative to them, and these experiments do not establish what masked reconstruction would contribute from a random start.

This choice mattered more than any other examined. Compared with random initialisation, self-supervised weights raised mean IoU by 0.0949 when only the encoder was transferred, and by 0.1074 when both encoder and decoder were (both $p < 0.01$). These three arms used the same learning-rate settings, so this comparison isolates the effect of the initial weights.

Both self-supervised variants also beat ImageNet initialisation, by 0.0428 and 0.0553 mean IoU ($p < 0.01$). This comparison is less clean, because the ImageNet arm also used different encoder learning-rate multipliers: 0.50, 0.10, 0.10, 0.15 and 0.25 times the base rate for the stem and four stages, against 0.25, 0.40, 0.60, 0.80 and 1.00 in the other arms. The difference therefore reflects initialisation and learning rate together. As with focal depth, the largest per-structure gain was for



Supplementary Figure S1: Segmentation accuracy against the number of focal depths supplied to the model. Points show the mean across ten patient-level cross-validation splits and bars the standard deviation; asterisks denote a significant difference from the full 11-depth model (two-sided paired Wilcoxon signed-rank test, * $p < 0.05$, ** $p < 0.01$). Mean IoU is over all five classes, background included.

the RB (Dice 0.18 to 0.24).

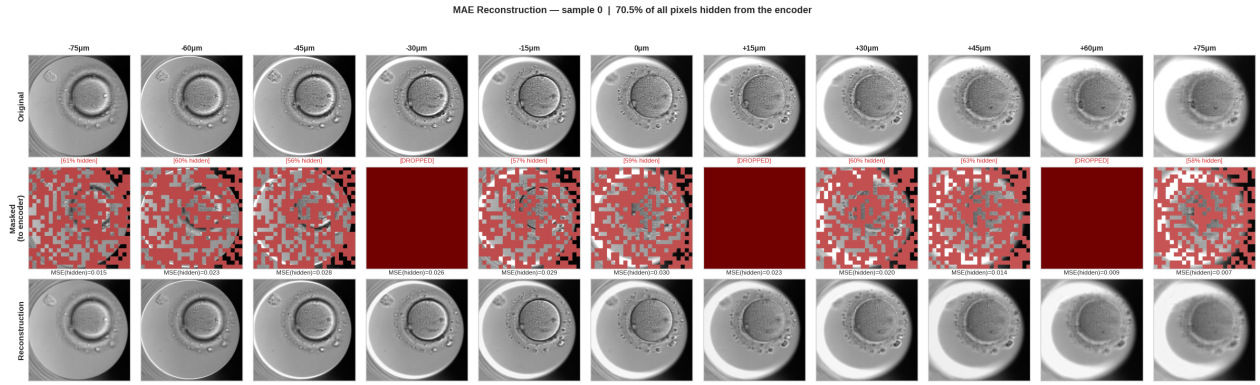
S3.4 How evidence from different depths is combined

Two simpler alternatives to the learned per-depth weighting were tested: averaging all depths equally, and using only the single depth the model judged best. Using a single depth was significantly worse (-0.0117 mean IoU, $p < 0.05$), even though all eleven depths remained available to the model as input: committing to the single most visible depth at the aggregation stage discards evidence that the remaining depths still carry. Against equal averaging the difference was small and not statistically significant (-0.0035), though accuracy for the RB was lower under equal averaging (Dice 0.14 vs. 0.18). The advantage of learning the weights rather than fixing them is therefore suggestive rather than established, and rests on the structure whose measurements vary most between splits.

The claim made in the main text follows from this. It is not that learned weighting improves segmentation, but that it provides a per-structure focal-depth profile, which equal averaging cannot, alongside a segmentation of comparable accuracy.

S3.5 Size of the image-analysis network

Three network sizes were compared. The largest (ResNet50) reached 0.6213 ± 0.0104 mean IoU, exceeding the reference by 0.0196 ($p < 0.01$); the intermediate size (ResNet34) gave a smaller improvement that was not significant ($+0.0100$). The gain from the largest network was almost entirely in the PVS (Dice 0.73 vs. 0.67). At this dataset size, additional network capacity was therefore mildly beneficial rather than harmful.



Supplementary Figure S2: Self-supervised initialisation by masked reconstruction of the focal stack. Top: the eleven acquired depths of one unannotated oocyte. Middle: the masked input, with hidden regions in red — three of the eleven depths masked in their entirety and 60% of 32×32 patches masked on each remaining depth, hiding 70.5% of image content in this example. Bottom: the reconstruction, produced without any use of annotations. Plausible content is recovered even at the wholly masked depths, indicating that the network draws on structure shared across the stack rather than treating each depth independently. Mean squared reconstruction error over the hidden regions, from -75 to $+75\,\mu$m: 0.015, 0.023, 0.028, 0.026, 0.029, 0.030, 0.023, 0.020, 0.014, 0.009 and 0.007. These are computed on channel-standardised intensities (Supplementary Table S6), whose per-plane standard deviation rises by half across the stack, so the apparent decline toward positive depths partly reflects that scaling.

S3.6 Order in which overlapping structures are resolved

Where two structures are both predicted at the same pixel, the model's cascading normalisation must award the pixel to one of them. The reference configuration does this by anatomical containment: RB before PB, PB before PVS, PVS before cell body. To test whether this order matters, it was reversed. The annotation was left unchanged, so every arm was scored against the same ground-truth masks and only the model's resolution order differed.

Accuracy for the PB and RB was essentially unchanged ($\Delta + 0.0005$ and $\Delta - 0.0069$ Dice, neither significant). The only significant change was a small gain for the cell body ($+0.0049$, $p < 0.01$). This is expected: with the cell body first, a pixel that the model considers plausibly either cell body or a smaller structure now goes to the cell body.

RB accuracy varies widely between splits (Dice 0.1803 ± 0.1011), so this comparison has little power for the structure most affected by overlap. The resolution order therefore does not appear to be a sensitive choice at this sample size, but that does not show that anatomical containment carries no useful information.

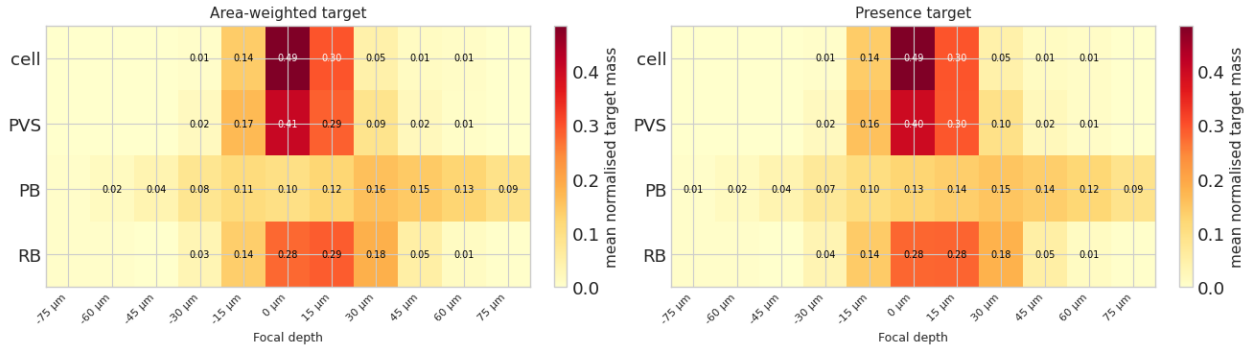
S3.7 How the focal-depth target is defined

Two ways of building the focal-depth training target from the annotations were compared. Under *presence* weighting, every depth at which a structure was annotated is given equal weight; under *area* weighting, each depth is additionally weighted by the number of pixels the structure occupies there, with a floor of 10% of the structure's largest annotated area applied to depths at which it is present. The two gave indistinguishable segmentation accuracy (-0.0014 mean IoU, not significant, with no per-structure differences).

This is not because the two targets are the same. They coincide only when a structure is

annotated at a single depth. That is almost always the case for the cell body (99.9% of development-set oocytes), but not for the other structures: the two targets differ for 69% of oocytes with an annotated PVS, 85% with a PB and 45% with an RB (Supplementary Table S7). Three of the four structures therefore received different supervision in the two arms, yet segmentation accuracy was the same.

One likely reason is that the cell body, which makes up 82.5% of annotated area and so dominates the class mean, is the one structure whose target does not change. We did not test whether the choice of target changes the predicted focal-depth profiles themselves, as opposed to segmentation accuracy. Averaged over oocytes, the two targets differ appreciably only for the polar body (Supplementary Fig. S3), the one structure annotated across most of the stack.



Supplementary Figure S3: The two focal-depth target constructions compared. Each panel gives the mean normalised target mass per structure and depth, averaged over the oocytes in which that structure is annotated: area-weighted (left) and presence (right). The rows for the cell body are identical at the precision shown, and those for the perivitelline space and refractile body differ by at most 0.01; only the polar body, the one structure annotated across most of the stack, differs appreciably. These are population means: the per-oocyte targets differ far more often (Supplementary Table S7).

S4 Data splits, preprocessing constants and annotation statistics

The 897 oocytes came from 208 patients: 778 oocytes from 187 patients in the development set and 119 oocytes from 21 patients in the held-out test set. All oocytes from a given patient were assigned to the same side of the split, and the split was stratified by oocyte quality grade. 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), although grade 3 was somewhat more frequent in the test set (47.9% against 38.9%). The ten cross-validation splits used in Supplementary Note S3 were also patient-level. In the structures annotated, the two sets are closely matched, the refractile body being somewhat more frequent in the test set (34.5% against 28.9% of oocytes).

Supplementary Table S6 gives the per-focal-plane constants used to standardise every stack. The plane mean rises monotonically from 0.504 at $-75\,\mu$m to 0.588 at $+75\,\mu$m and the standard deviation from 0.201 to 0.306, a systematic optical gradient across the stack rather than sample-to-sample variation. The planes of a stack are therefore not interchangeable in intensity or contrast, which is relevant both to the reconstruction errors quoted in Supplementary Fig. S2 and to any method that collapses the stack by projection.

Supplementary Table S7 reports how many focal depths each structure's annotation spans. The

Supplementary Table S4: Development-set comparison across ten patient-level cross-validation splits. The first block gives absolute mean IoU for the reference (mean ± standard deviation) and the selected configuration; all remaining rows give the change in mean IoU relative to the reference. Mean IoU is over all five classes, background included. Significance of each difference (two-sided paired Wilcoxon signed-rank tests, $n = 10$, exploratory) is reported in the text of Supplementary Note S3.

The reference is ImageNet-initialised and differs from the selected configuration in initialisation; the initialisation comparison against ImageNet additionally differed in learning-rate settings (Supplementary Note S3.3). The selected configuration is the same arm as "from oocyte images, both halves" and is repeated in the first block for convenience.

Component varied	**Variant**	**Mean IoU**
—	Reference configuration (ImageNet)	0.6017 ± 0.0116
—	**Selected configuration** (oocyte images, both halves)	0.6570
		Δ vs. reference
Combining depths	Equal averaging	−0.0035
	Single depth only	−0.0117
Overlap resolution	Reversed order	−0.0011
Focal-depth target	Presence weighting	−0.0014
Initial weights	Random	−0.0521
	From oocyte images, encoder	+0.0428
	From oocyte images, both halves	+0.0553
Network size	ResNet34	+0.0100
	ResNet50	+0.0196
Number of focal depths	1 (central)	−0.0373
	3	−0.0346
	5	−0.0172
	7	−0.0105

cell body is annotated at a single depth in essentially every oocyte, whereas the PVS and PB span two and approximately three depths on average; the RB, though rare, is annotated at a single depth in more than half of the oocytes in which it appears. These statistics govern how the presence and area target constructions relate to one another (Supplementary Note S3.7) and how closely a predicted distribution can in principle match the annotation for each structure, since a target concentrated on one depth penalises a diffuse prediction more heavily than one spread over three.

Supplementary Table S8 gives the mean area-weighted target for each structure and depth. The cell body and PVS are concentrated at 0 and +15 µm, which together carry 78% and 69% of their mass in the development set (78% and 66% in the test set). The polar body is the exception: its population profile is nearly flat, spanning every depth from −75 to +75 µm with no value above 0.16 in the development set, while individual oocytes are annotated at 2.7 to 3.0 depths on average (Supplementary Table S7). The depths at which a polar body is assessable therefore vary from oocyte to oocyte, and the population profile is a poor guide to them. This is the situation the framework is designed for. The development-set and test-set profiles are closely similar, indicating that the patient-level split did not separate oocytes with systematically different focal behaviour.

Chance level of the PR-AUC For each oocyte and structure, PR-AUC (computed as average precision) asks whether the depths at which the structure was annotated are ranked above the depths

Supplementary Table S5: Composition of the development and held-out test sets. Top: number of oocytes in which each structure was annotated. Bottom: oocyte quality grade, by which the split was stratified.

Percentages are of the oocytes in that split.

	Development	**Test**	**Total**
Patients	187	21	208
Oocytes	778	119	897
Structure annotated			
Cell body	775 (99.6%)	119 (100%)	894
Perivitelline space	747 (96.0%)	115 (96.6%)	862
Polar body	741 (95.2%)	117 (98.3%)	858
Refractile body	225 (28.9%)	41 (34.5%)	266
Quality grade			
0	190 (24.4%)	27 (22.7%)	217
1	168 (21.6%)	18 (15.1%)	186
2	117 (15.0%)	17 (14.3%)	134
3	303 (38.9%)	57 (47.9%)	360

at which it was not. Its value under a random ranking is not the fraction of annotated depths, as it would be for a precision–recall curve over many items, because with only eleven depths the first positive is often reached early. If k of the 11 depths are annotated and the ranking is random, the annotated depths occupy ranks $r_1 < \cdots < r_k$ drawn uniformly without replacement, and

$$\mathbb{E}[\mathrm{AP}] \;=\; \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[\frac{i}{r_i}\right], \tag{S3}$$

since the precision at the i-th annotated depth is i/r_i. We evaluated (S3) exactly by enumerating all $\binom{11}{k}$ placements; a simulation with 20,000 random rankings per k agreed to two decimal places. The results are in Supplementary Table S9. Since structures are annotated at between one and three depths on average (Supplementary Table S7), the relevant chance levels lie between about 0.27 and 0.42, well above the $1/11 \approx 0.09$ that the prevalence of a single depth would suggest. Rows in which all or none of the depths are annotated were excluded from the PR-AUC, as average precision is undefined or trivial for them.

S5 Training curves

This section shows how the selected model's performance evolved over training: overall segmentation (Supplementary Fig. S4), segmentation by structure (Supplementary Fig. S5) and focal-depth prediction (Supplementary Fig. S6). The per-depth breakdown of focal-depth prediction, and a worked example of predicted structures overlaid on the full stack, are in the main text.

Note on the curves that follow The curves show held-out test-set metrics at each epoch of the final training run, with training-set metrics where indicated. The architecture, hyper-parameters and number of epochs were all fixed by the development-set cross-validation of Supplementary Note S3 before this run began. The main text reports the values at the final epoch. No checkpoint, epoch or

Supplementary Table S6: Per-focal-plane normalisation constants. Each plane of every stack was centred and scaled by the corresponding values, computed over the 778 development-set stacks on $[0, 1]$-scaled grayscale intensities.

Focal depth (μm)	**Mean**	**s.d.**
−75	0.5035	0.2005
−60	0.5133	0.2115
−45	0.5213	0.2259
−30	0.5301	0.2401
−15	0.5392	0.2522
0	0.5484	0.2628
+15	0.5571	0.2727
+30	0.5653	0.2820
+45	0.5733	0.2907
+60	0.5808	0.2986
+75	0.5881	0.3062

Supplementary Table S7: Span of the focal-depth annotations. "Mean depths" is the mean number of depths at which a structure is annotated, over the oocytes in which it is annotated at all; "single depth" is the proportion of those oocytes annotated at exactly one depth; "identical targets" is the proportion for which the presence and area constructions normalise to the same target row, and which therefore receive identical supervision in the two arms of Supplementary Note S3.7.

	Development set			**Test set**		
Structure	**Mean depths**	**Single**	**Identical**	**Mean depths**	**Single**	**Identical**
Cell body	1.00	99.9%	99.9%	1.02	99.2%	99.2%
Perivitelline space	1.97	30.9%	30.9%	2.10	27.0%	27.0%
Polar body	2.67	14.8%	15.0%	2.99	6.8%	7.7%
Refractile body	1.68	55.1%	55.1%	1.44	61.0%	63.4%

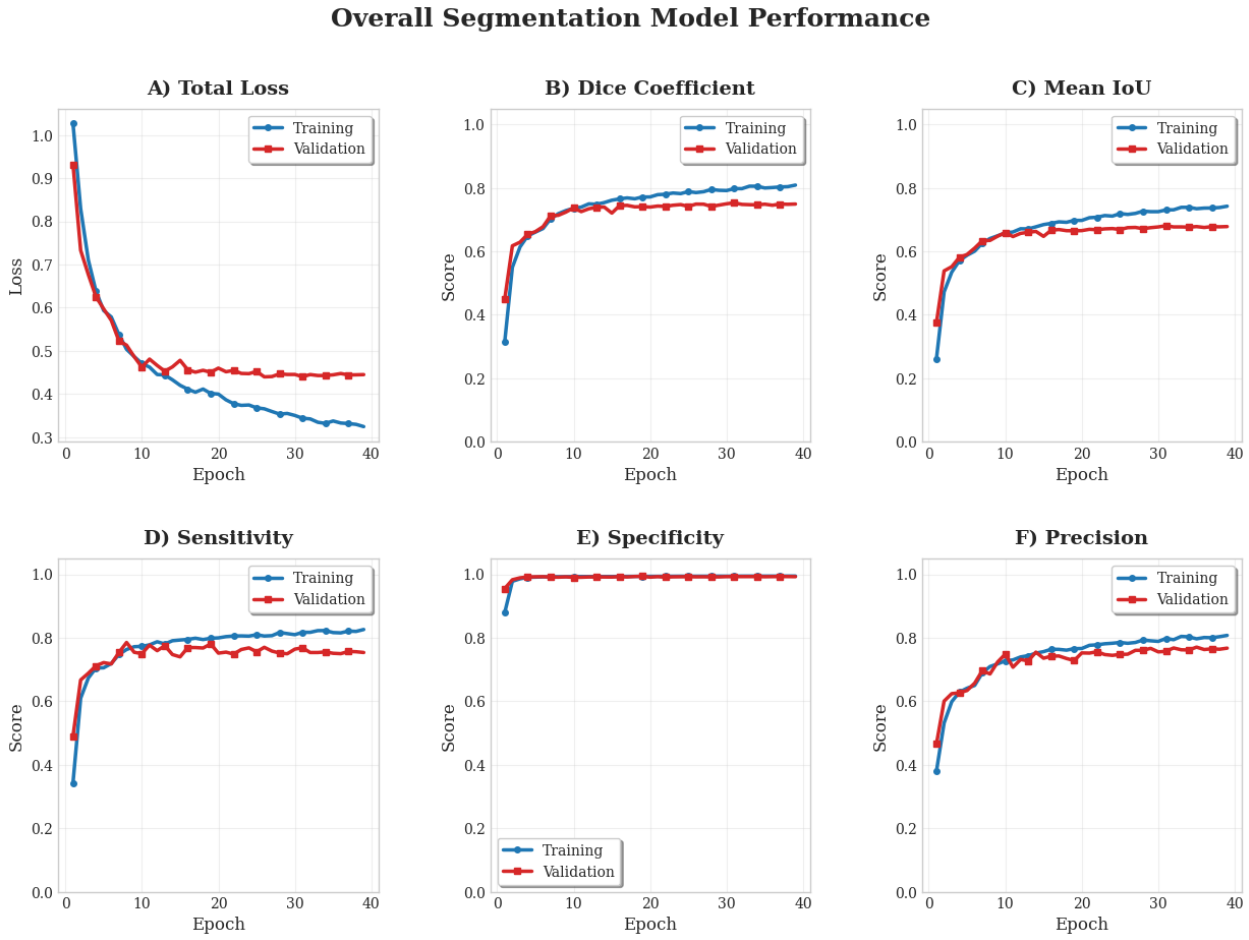
design choice was selected using these curves; they are shown so that the course of training can be inspected.

Supplementary Table S8: Mean normalised area-weighted focal-depth target by structure and depth, averaged over the oocytes in which each structure is annotated. Columns sum to one up to rounding.

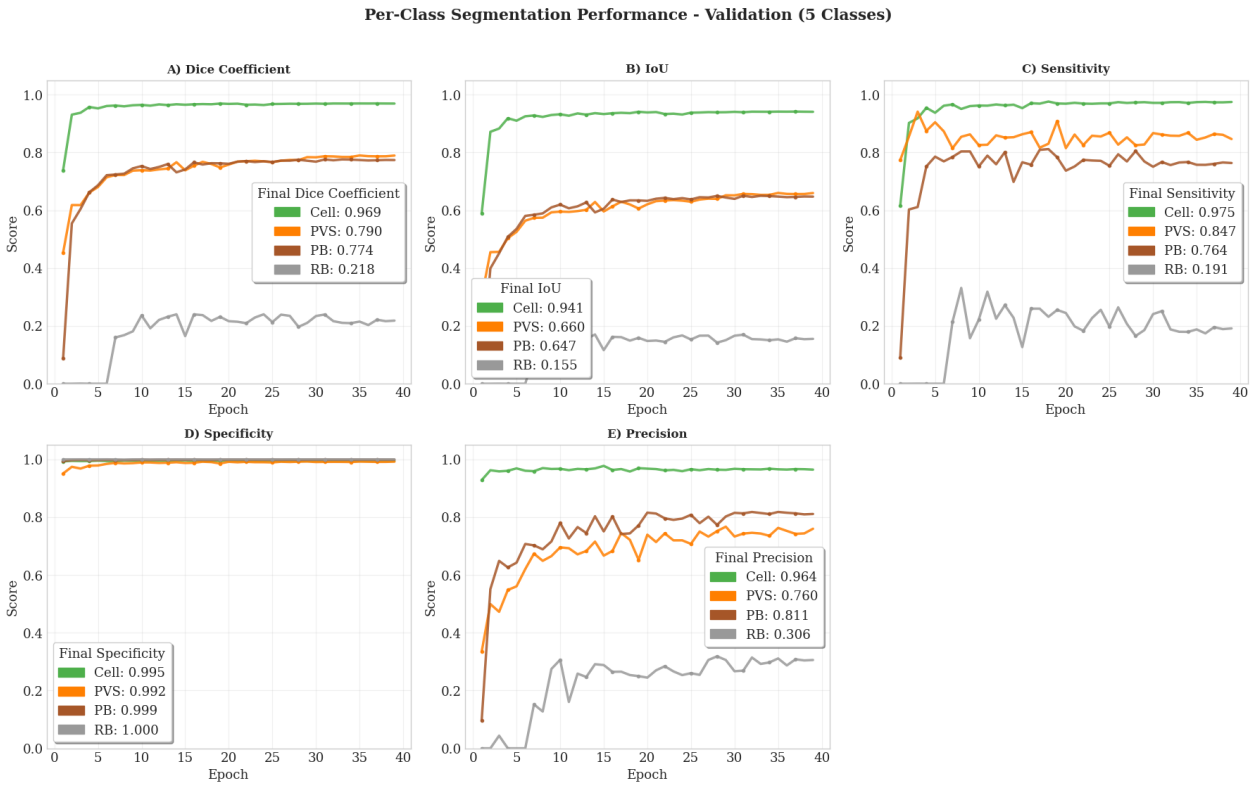
	Development set				Test set			
Depth (μm)	**Cell**	**PVS**	**PB**	**RB**	**Cell**	**PVS**	**PB**	**RB**
−75	0.000	0.000	0.005	0.004	0.000	0.000	0.004	0.000
−60	0.000	0.000	0.017	0.004	0.000	0.000	0.009	0.000
−45	0.001	0.000	0.037	0.001	0.000	0.000	0.016	0.000
−30	0.005	0.019	0.082	0.033	0.000	0.025	0.050	0.066
−15	0.142	0.170	0.111	0.137	0.101	0.149	0.111	0.128
0	0.485	0.407	0.099	0.277	0.412	0.372	0.097	0.207
+15	0.297	0.285	0.125	0.289	0.370	0.283	0.111	0.234
+30	0.047	0.088	0.160	0.184	0.067	0.116	0.199	0.207
+45	0.014	0.023	0.145	0.054	0.045	0.049	0.154	0.106
+60	0.008	0.007	0.126	0.013	0.003	0.006	0.146	0.029
+75	0.001	0.001	0.093	0.003	0.003	0.000	0.103	0.024

Supplementary Table S9: Expected PR-AUC (average precision) of a random ranking of the 11 focal depths, by the number k of depths annotated. Computed exactly from (S3).

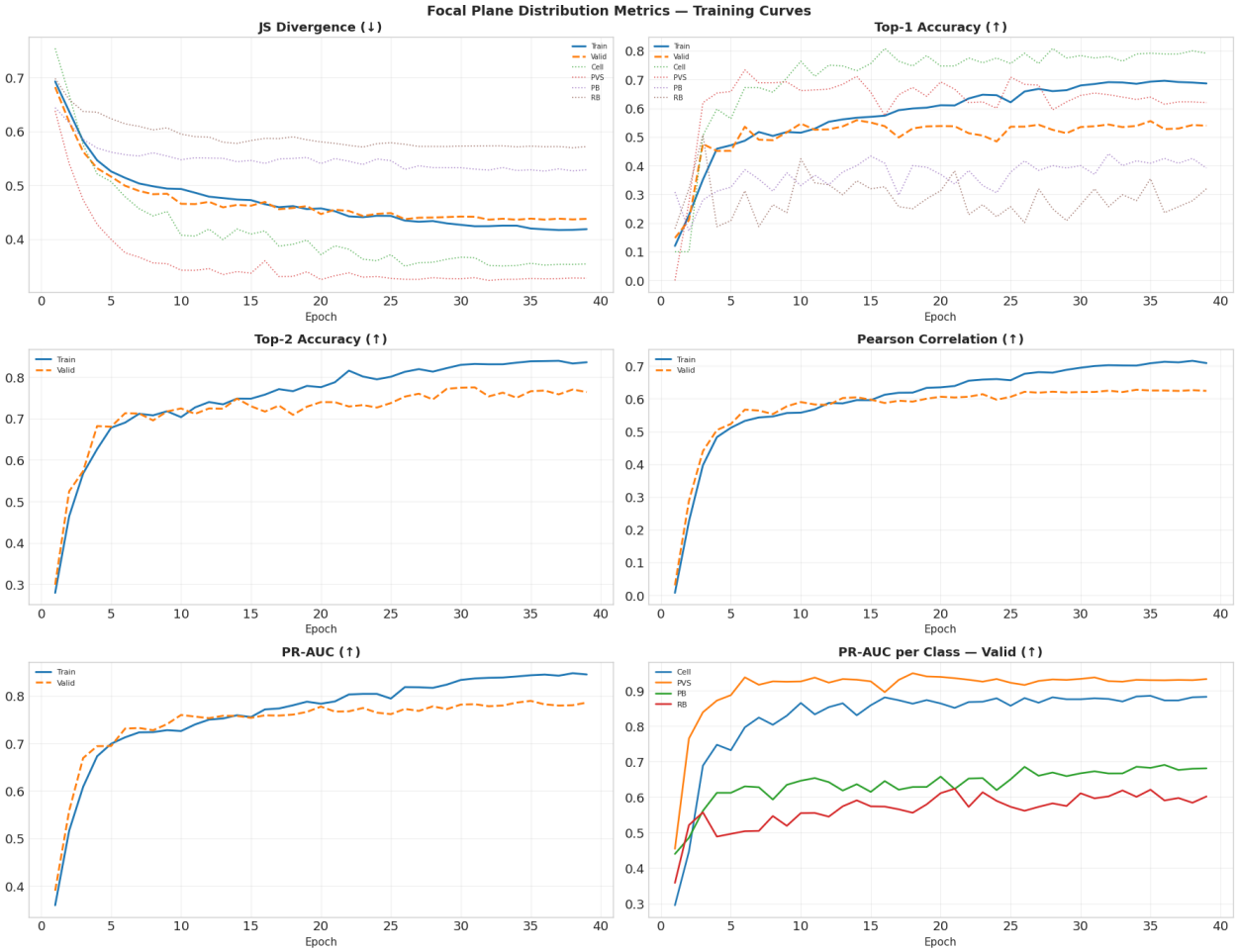
Annotated depths k	**Expected PR-AUC**
1	0.275
2	0.347
3	0.420
4	0.492
5	0.565



Supplementary Figure S4: Course of training for the selected model. Total loss (A), Dice coefficient (B), mean intersection-over-union (C), sensitivity (D), specificity (E) and precision (F) are shown for the training set and the held-out test set at each epoch. Means are taken over all five classes, background included. Final held-out values: Dice 0.749, mean IoU 0.678, against 0.81 and 0.74 on the training set.



Supplementary Figure S5: Accuracy for each of the four annotated structures over the course of training on the held-out test set: Dice coefficient (A), intersection-over-union (B), sensitivity (C), specificity (D) and precision (E). Final values are given in the legend of each panel and are reproduced in Table 1 of the main text. The refractile body is separated from the other three structures by a wide margin on every measure except specificity, which is uninformative here because background dominates.



Supplementary Figure S6: Focal-depth prediction over the course of training, for the training set and the held-out test set. In the top two panels, dotted lines show the four structures separately (test set). **Top left**: Jensen–Shannon distance between the predicted and annotated depth profiles (lower is better; maximum $\sqrt{\ln 2} \approx 0.83$). **Top right**: peak overlap. A structure's *peak depths* are those within 90% of the profile's maximum; peak overlap is the proportion of cases in which the predicted and annotated peak depths share at least one depth. **Middle left**: two-depth overlap, the proportion of cases in which the two highest-weighted predicted depths include an annotated peak depth. **Middle right**: Pearson correlation between predicted and annotated profiles across the eleven depths. **Bottom left**: PR-AUC, which asks whether annotated depths are ranked above unannotated ones and needs no threshold; a random ranking scores about 0.27 to 0.42, depending on how many depths are annotated (Supplementary Table S9). **Bottom right**: PR-AUC by structure, held-out test set. Final held-out values: Jensen–Shannon distance 0.437, peak overlap 53.5%, two-depth overlap 77%, Pearson correlation 0.62, PR-AUC 0.79. Peak overlap shows the largest gap between training and held-out performance (70% against 53.5%); the Jensen–Shannon distance shows almost none (0.42 against 0.44). The two measures rank the cell body and PVS in opposite order because the cell body is annotated at a single depth in 99% of oocytes (Supplementary Table S7). A single-depth target is easy to hit at the peak, but any spread in the prediction counts heavily against it as a distribution.