

FollicleFinder enables automated three-dimensional segmentation of human ovarian follicles

Leopold Franz · Kevin A. Yamauchi · Selina Fricke · Marieke Biniash · Harold Gómez · Christian De Geyter · Dagmar Iber

Received: 25 April 2023 / Accepted: 29 August 2026

© The Author(s) 2026

Abstract

In vitro fertilization (IVF) treatment protocols require frequent monitoring of the ovarian follicle growth process. Current treatment protocols are financially, physically, and psychologically burdensome for the patient. Mechanistic models of folliculogenesis have the potential to improve outcomes by personalizing treatments to the individual patient, but require high-fidelity data of reliable quality. We report FollicleFinder, an open source pipeline for automated, 3D segmentation and measurement of ovarian follicles in ultrasound images. FollicleFinder was developed using local shape descriptors as an auxiliary task to improve segmentation accuracy. It was validated using expert-labeled ground truth data. To make FollicleFinder amenable for use by clinicians, we created a graphical user interface for performing the segmentation and interactively exploring the results. Further, we are releasing our 533 sets of raw images and expert-labeled ground truth as a benchmark dataset. Together, the open source segmentation pipeline, graphical user interface, and benchmark dataset provide a suite of tools for accurate 3D follicle segmentation and measurement instrumental for more predictive personalized IVF treatment models capable of guiding critical clinical decisions, such as dosage adjustments and trigger timing.

Introduction

Assisted reproductive technology (ART) offers a range of treatment options to couples having trouble conceiving. In vitro fertilization (IVF) treatment protocols include the collection of gametes. IVF commonly requires a set of multiple oocytes and ovarian stimulation is therefore used to support the maturation of a large number of ovarian follicles. During the hormone stimulation, the ovarian response is monitored by repeated transvaginal ultrasound (TVUS) examinations, and clinical decisions regarding the dosage and timing of hormonal stimulation are based on the follicle number and size. Current IVF treatment imposes significant financial, physical, and psychological burden on patients owing to the frequent doctor visits and the hormonal treatment^{1,2}.

The current IVF live birth rate often requires patients to undergo multiple rounds of treatment. Recent advances in endocrinology and mathematical modeling have yielded mechanistic models that predict follicle growth

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.

Leopold Franz, Kevin A. Yamauchi, Selina Fricke and Marieke Biniash contributed equally to this work.



dynamics from hormonal concentrations and manual 2D follicle diameter measurements^{3–7}. These models are an important step towards improving IVF treatment by model-driven personalized treatment. Using such models, one could optimize the hormone stimulation protocol for a given patient, decreasing unnecessary high hormone dosages and their side-effects, reducing the doctors time per patient and increasing the success rate^{6,8–11}. While promising, in order to implement and validate these models in the clinic, there is a need for software that allows researchers to integrate 3D follicular measurements with models in a clinical setting.

Previously, only 2D slices were imaged in clinical settings, but 3D ultrasound scanners and 3D segmentation algorithms have since been developed¹². Unlike traditional 2D approaches, automated 3D measurements eliminate operator dependency and prevent the over- or underestimation of size in less spherical follicles, which are discrepancies that can lead to poor clinical decisions under current size-dependent protocols. Several companies now offer commercial software packages for the 3D segmentation of ovarian follicles¹². General Electric (GE) offers two 3D segmentation algorithms, VOCAL and SonoAVC¹³, as add-ons to its ultrasound scanner. Both software packages measure the volume and surface area of follicles¹⁴. Unfortunately, the prevalence of segmentation errors requiring manual correction hinders the routine clinical adoption of these packages¹⁵. Measurements by hand introduce an operator-dependent variability¹⁶. Furthermore, proprietary systems act as ‘black boxes’ that withhold full phenotypic data, such as exact follicle geometry and spatial distribution, thereby obstructing integration with comprehensive computational models. Accordingly, there is a need for improved algorithms that enable the automatic, accurate detection and 3D segmentation of ovarian follicles in a modular package that can connect the 3D follicle morphometric measurements with computational models to guide patient treatments.

Deep learning approaches have improved image segmentation in biological images^{17,18}, and have recently been explored in the context of ovarian follicle detection and segmentation. CR-UNet, which combines a spatial Recurrent Neural Network (RNN) with a standard 2D U-Net, yields promising results with 2D TVUS images^{12,16}. While promising, the CR-UNet has lower accuracy with large follicles, which are clinically important as the largest follicle is generally the ovulating follicle. A subsequent study used an S-Net architecture for the simultaneous segmentation of ovary and follicles in 3D TVUS volumes¹⁹. More recently, Potočník et al.²⁰ introduced ‘USOVA3D’, the first public database of annotated 3D ovarian ultrasound volumes and evaluated a standard 3D U-Net along with deeply-supervised variants²¹. Additionally, a weakly-supervised framework has been proposed to label 3D follicles by segmenting 2D slices using ‘USOVA3D’ and an alternative dataset of 2D ultrasound images called the ‘FUID’ dataset²². While these diverse deep learning and database developments represent valuable milestones in automated follicle tracking, they focus on dataset releases under highly restrictive licenses or architectural evaluations without accompanying open-source model weights, or end-user graphical interfaces. Consequently, a critical gap remains for fully reproducible, open-source 3D tools that clinicians and researchers can readily deploy, adapt, and build upon to improve clinical outcomes.

To facilitate the translation of 3D image-derived features into the clinic and the development of new models of follicular selection, there is a need for open-source software that will serve as the basis for these innovations. To fill this gap, we have developed FollicleFinder, an open source platform for the detection and measurement of follicles from transvaginal ultrasound images and FollicleFinder a graphical user interface for segmentation and morphometric analysis of ovarian follicles from transvaginal ultrasound images. As the complete segmentations are directly available at the end of the pipeline new downstream analysis can easily be added. Finally, we have produced and made available the largest open dataset of 533 transvaginal ultrasound ovarian images with expert-annotated labels for ovary, follicles.

Methods

Approach

FollicleFinder (Fig. 1B) performs instance segmentation of the follicles from transvaginal ultrasound images using a 3D Resnet and a 3D UNet via a two stage segmentation pipeline (Fig. 1A). Instance segmentation is a technique that identifies and separates individual objects (follicles) as distinct entities rather than classifying them into a single class. First, we predict local shape descriptors from the raw image²³. Next, concatenate the raw image with the predicted local shape descriptors and predict the follicles. We propose predict local shape descriptors as an intermediate task as they have been shown to improve boundary predictions and separation of adjacent instances²³. Morphological properties such as the volume, surface area, and effective diameter are directly measured from the segmented follicles. Thus, follicle population growth dynamics can be tracked over the course of the menstrual cycle (Fig. 1C).

To facilitate usage of the FollicleFinder pipeline in research and clinical settings, we have created a graphical user interface as a napari plugin (Fig. 1B, Figure S1). The FollicleFinder plugin loads images from standard image formats (e.g., TIFF, DICOM, ZARR, HDF5) and renders them in 2D and 3D in the napari viewer. Via the graphical user interface, the user can use FollicleFinder to segment the follicles and measure their properties. Since FollicleFinder is written as a napari plugin with standard python libraries, we expect that others will be able to use and extend it with other analyses.

Ethical statement

The study was presented to and approved by the local ethics review board (BICYCLE Study: EKNZ ID 2016-01824, Prospect Human Study: EKNZ ID 2022-01969). All activities were conducted in accordance with the Declaration of Helsinki of 1975 (revised 2013), the International Council on Harmonisation of Technical

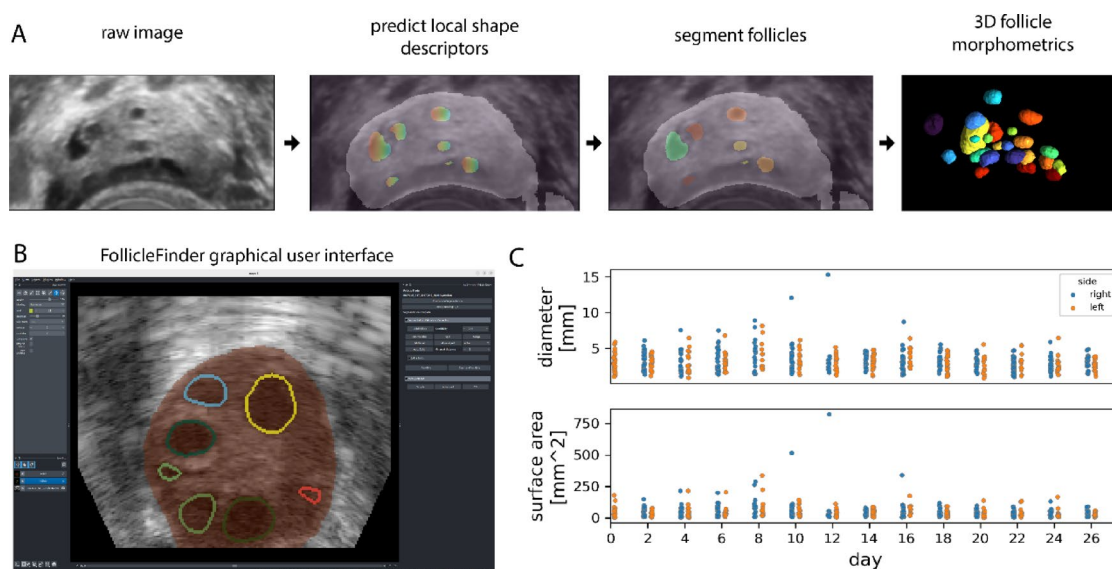


Fig. 1 (A) Overview of the processing pipeline in which the local shape descriptors are predicted from the raw image, the follicles are segmented, and morphological measurements are made. (B) The FollicleFinder pipeline can be run via the command line or via the FollicleFinder napari plugin graphical user interface (GUI). With the GUI, users can perform the segmentation and then interactively view the resulting geometries and measurements. (C) Example of the longitudinal morphometry of ovarian follicles when segmenting ultrasound images of Patient A17 Cycle 2 images with the fold 0 model. Note the emergence of the dominant follicle in the right ovary on day 10.

Requirements for Registration of Pharmaceuticals for Human Use Good Clinical Practice guidelines and current national guidelines. All participants provided written informed consent.

Image acquisition

We use transvaginal ultrasound images of both ovaries from 29 women between 21 and 39 years old, that participated in the BICYCLE study²⁴. To compare the analysis time with FollicleFinder and SonoAVC, we measured the time to complete the 3D image acquisition as well as the time to complete the 3D image segmentation and analysis versus the time to complete all 2D measurements, which includes 5 follicle measurements per ovary and one endometrial thickness measurement. We conducted this timing analysis with 6 different doctors, as part of the PROSPECT study.

The imaging as part of the BICYCLE study was carried out using a GE Healthcare Voluson E8 ultrasound machine and RIC5-9-D ultrasound probe. For the timing analysis and the comparison to SonoAVC we used GE's Voluson Signature 20 ultrasound machine and RIC 5-9-D ultrasound probe with the latest version of SonoAVC¹³ on April 6th 2026.

Image export and ground-truth dataset generation

The 3D images were exported in the DICOM format using GE 4D View software and converted to tiff images. For initial training and testing, a clinical expert hand-segmented 98 ovaries using the separate 3D slicer program²⁵. If the expert was unsure one other expert was included in the process to find an agreement. For a subset of the segmentations a second clinical expert reviewed the segmentations and corrected obvious mistakes (merges, splits, clearly missing follicle sections). All other mistakes were again discussed with the first clinician to find a consensus. For six ovary volumes from the first ultrasound segmentation dataset, three experts independently segmented all follicles from scratch using the napari plugin²⁶ without access to each other's annotations or an initial model output. The remaining 385 validated segmentations were corrected by a clinician or an ultrasound expert starting from the segmentation results of the first version of the model trained on the initial 98 hand segmented images or based on later interim versions. Similar to the initial labelling process, another clinician was involved in the process if the expert was unsure.

Pre-processing

To prepare the ultrasound images for processing, we first rescaled the images to $0.157288 \text{ mm} \times 0.157288 \text{ mm} \times 0.157288 \text{ mm}$ voxels. We then padded images, so they have a minimal length of 80 along each axis as both models had a context chunk size of $80 \times 80 \times 80$ voxels. We then replaced the padding voxels to the mean ultrasound intensity and applied some Gaussian blurring to the transition towards the padding voxels. The padding voxels are the voxels added to yield a rectangular image from cone shaped detection volume. Next, we applied Contrast Limited Adaptive Histogram Equalization (CLAHE) to the image to improve the local contrast of the different regions. After we added a Gaussian blurring with $\sigma = 10$ on the image and converted the image back to uint8. During training the images were then sliced into $80 \times 80 \times 80$ chunks, and a final z-score standardization was applied to each chunk.

To prepare the labels for the second model we calculated the combined inner and outer follicle boundaries and expanded them by 1 voxel. We then combined the ovary, follicle and follicle boundary masks into a single multiclass label array.

Model architecture

We predicted the local shape descriptors using a 3D Resnet^{27,28}. The Resnet had a single input size channel corresponding to the pre-processed image. We used a feature map of (16, 32, 64, 128, 256) with 8 different groups. The model predicted the four local shape descriptors as described in “[Local shape descriptors](#)” section. For the follicle segmentation we used a 3D UNet^{28,29}. The 3D UNet had five input channels corresponding to the four predicted local shape descriptors and the pre-processed image. The model predicted the four output classes: background, ovary, follicle and follicle boundary. We used a feature map of (32, 64, 128, 256) with a single group. We used both models as implemented by Wolny et al.²⁸.

Post-processing

To post-process the follicle predictions we used a 0.5 threshold to extract the follicle voxels. We then identified individual follicles via connected component analysis. Next, we expanded each follicle by one voxel while keeping the individual follicles separate by at least one voxel.

Local shape descriptors

To improve the segmentation of follicles, we implemented an auxiliary task to predict local shape descriptors. Sheridan et al.²³ used local shape descriptors to improve neuron segmentation performance in volume electron microscopy images. Shape descriptors provide local information that helps separate proximal follicles and reduce merge errors. Following their work, we implemented shape descriptors for the vector pointing to the centroid of the follicle. We also implemented a new shape descriptor, which is the inverse squared distance of each ovary pixel to the nearest follicle boundary with a per-image 98th-percentile normalization applied to the distance (Fig. 2).

Training local shape descriptors

We trained the local shape descriptor neural network using the ADAM optimizer. We set the initial learning rate to 2.5E-4 and the weight decay to 1E-5. We reduced the learning rate by multiplying a factor of 0.2 after no improvement over 10 validation runs. We used a weighted mean squared error loss defined in Eq. (1).

$$L_{MSE} = \frac{\sum_{n=1}^N w * (x_n - y_n)^2}{N} \quad (1)$$

x_n and y_n are the voxel-wise prediction and ground truth respectively and N is the total number of voxels. The weight w is defined as in Eq. (2).

$$w = \begin{cases} c & \text{if } x < t \\ 1 & \text{if } x \geq t \end{cases} \quad (2)$$

c is a constant weight applied when the given voxel is less than the threshold t . We set $c = 0.25$ and $t = 0.005$.

We set the batch size to 30 and validated after 1000 iterations. We used the weighted mean squared error loss defined in Eq. (1) as our validation metric. To augment our training data, we applied random flips, random 90 degree rotations, random rotations (± 30 degrees), random elastic deformations, additive Gaussian noise (execution probability 0.2), and additive Poisson noise (execution probability 0.2). We stopped training when the learning rate was reduced to below 1E-6 or using an early stopping mechanism that stops training if after 24 validation steps the validation loss has not been improved by a minimum delta of 1E-4.

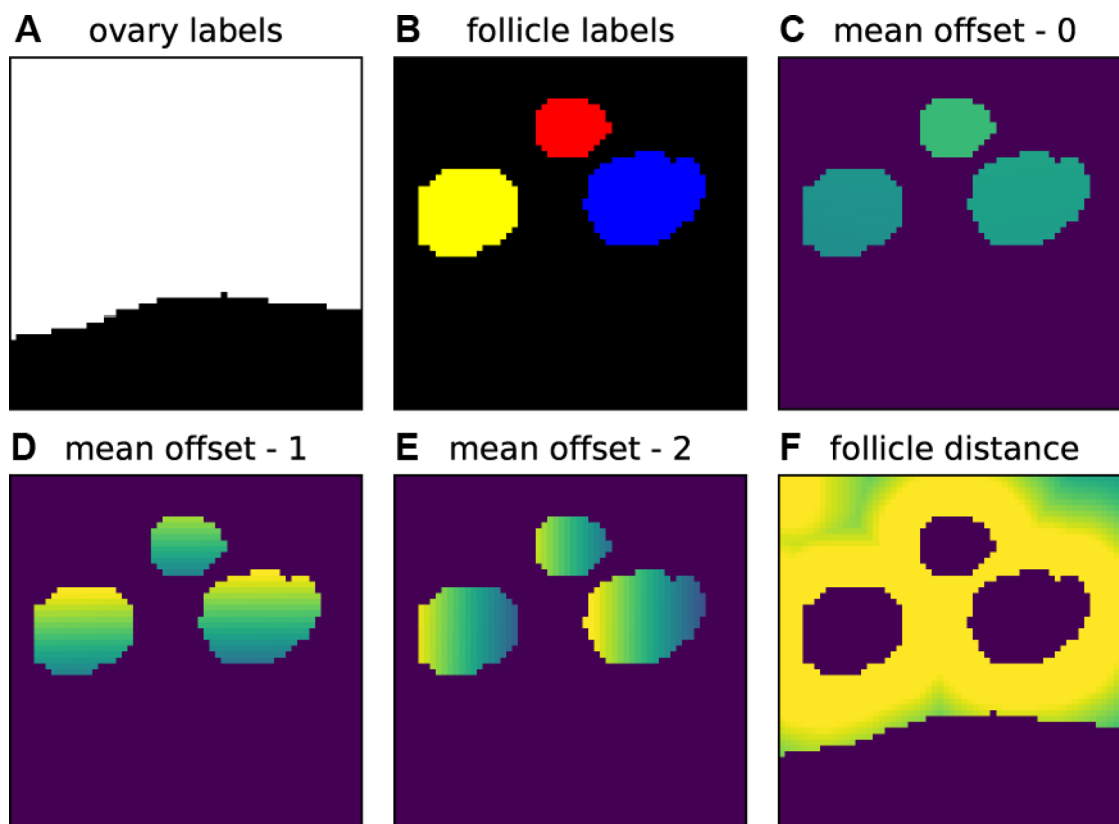


Fig. 2 Visualizing the local shape descriptor spatial features used to identify follicles. To help the AI accurately separate individual follicles, we provide it with specific “shape descriptors.” (A) A mask showing the boundary of the ovary. (B) The target map where each individual follicle is assigned a unique color/ID. Panels (C–E) represent the relative position (offset) of each pixel from the center of its follicle along the x, y, and z axes; these gradients help the model distinguish where one follicle ends and another begins to reduce the number of merged follicles. (F) A distance map showing how far each part of the ovary is from the nearest follicle, providing the model with additional spatial context for segmentation.

Training follicle segmentation

We trained the follicle segmentation network using the ADAM optimizer. As an auxiliary task, the network predicted the ovary and follicle boundaries in addition to the follicles themselves. We set the initial learning rate to $2\text{E}-4$ and the weight decay to $1\text{E}-5$. We reduced the learning rate by multiplying a factor of 0.2 after no improvement over 10 validation runs. We used a weighted cross entropy loss as described in³⁰ and written below in Eq. (3).

$$L_{WCE} = -\frac{1}{N} \sum_{n=1}^N w_c y_n \log(x_n) + (1 - y_n) \log(1 - x_n) \quad (3)$$

x_n is the voxel-wise predictions corresponding to float vector with a float ranging between 0 and 1 for each class. y_n is the ground truth corresponding to a vector with a binary entry for each output class c . N is the total number of voxels. The class weight w_c is defined as

$$w_c = \frac{1 - \sum_n x_{n,c}}{\sum_n x_{n,c}} \quad (4)$$

where $x_{n,c}$ is the softmax-normalized voxel-wise prediction for class c . This loss function therefore includes two terms, where the second term handles the case where voxels correspond to the background ($y_n = 0$). We used intersection over union of the follicle and ovary predictions as our validation metric. We set the batch size to 15 and validated after 1000 iterations. To augment our training data, we applied random flips, random 90 degree rotations, random rotations (± 30 degrees), and random elastic deformations. We stopped training when the learning rate was reduced to below $1\text{E}-6$ or using an early stopping mechanism that stops training if after 24 validation steps the validation loss has not been improved by a minimum delta of $1\text{E}-4$.

Inference

For both the local-shape-descriptor and the follicle-segmentation networks, we predict on full volumes by sliding-window inference with a patch size of $80 \times 80 \times 80$ voxels and a stride of $20 \times 20 \times 20$ voxels. Before inference we apply a $16 \times 32 \times 32$ mirror padding to the input volume so that voxels at the volume edges still receive contributions from multiple overlapping patches; the mirror padding is removed from the final prediction map. Per-voxel patch scores from overlapping patches are averaged.

Volume, surface area, and diameter measurements

We measured the volume of the segmented follicles with the scikit-image³¹ `regionprops_table` function. To measure the surface area of the follicles, we computed the surface area of a triangular mesh fit to the surface of the segmented follicle. First, we generated a mesh using the scikit-image marching cubes algorithm. Then we cleaned up the mesh (i.e., made it watertight) using `pymeshfix`³². Next, we smoothed the mesh using the `trimesh`³³ `filter_taubin` function (a Laplacian filter). Finally, we calculated the surface area using the `Trimesh.area()` method.

We also calculated the effective diameter using the volume by assuming the follicle is perfect sphere using Eq. (5).

$$d = 2 \left(\frac{3}{4\pi} V \right)^{1/3} \quad (5)$$

where d is the effective diameter and V is the volume.

FollicleFinder graphical user interface

We developed the graphical user interface (GUI) as a plugin for napari²⁶, a GPU-accelerated nD data viewer. Within the plugin, users can explore the data in both 2D and 3D. Our plugin is structured in three parts: IO (i.e., Input–Output: Functions defining how data is loaded and saved.), GUI (Fig. 9) and Model (Fig. 3).

napari allows users to import 3D images in various formats by drag-and-dropping them into the GUI. This includes images saved in the standard DICOM format. Note, that some GE ultrasound machines have a feature to save 3D ultrasound images as "DICOM", but then do not save them using the universal DICOM standard but a proprietary image file format³⁴.

Through the controls on the napari plugin GUI, users can start the segmentation of ovarian ultrasound images and morphological feature analysis, such as volume, surface area, effective diameter, sphericity, follicle centroid, distance to ovary centroid, distance to ovary boundary of all detected follicles.

The model that processes and analyses the input ovary images is implemented separately from the GUI to ensure all variables, images and settings only have one state. The model computation is run in a separate thread than the GUI, so that the viewer remains responsive during segmentation and analysis. Any intermediate results are directly sent back to the user interface thread and added to the viewer. The intermediate results are the local shape descriptors, the follicle predictions, and the processed follicle segmentation. The follicle morphological measurements are then displayed as a table and a plot.

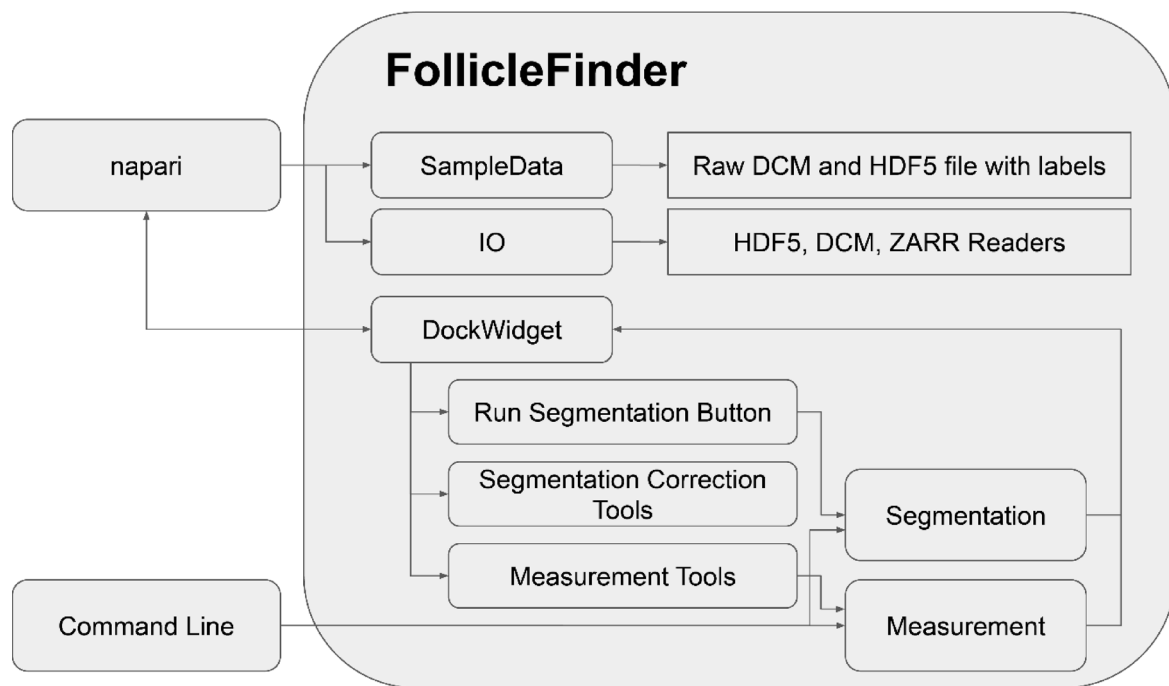


Fig. 3 FollicleFinder provides a graphical user interface for the ultrasound segmentation pipeline. FollicleFinder is implemented as a napari plugin and has a modular architecture. The modular architecture allows for rapid segmentation, segmentation correction and validation, and integration with mathematical models.

The FollicleFinder plugin allows users to save every 3D ultrasound image together with intermediate results and the final segmentations as hdf5/zarr/dcm files for each ultrasound image. The plugin can also export the measurements table as a csv file using the “Export Meas.” Button. The saved segmentations can then be reopened and edited in napari using the FollicleFinder plugin.

Comparing image acquisition and analysis times

To compare the clinical usability of FollicleFinder, we compared the time required for conventional 2D slice-based measurements with the time required for 3D volume acquisition of the uterus, left ovary, and right ovary. For the 3D approach, we recorded the time needed to acquire three high quality volumes immediately after probe insertion; when multiple attempts were required to obtain volumes of sufficient quality, this additional time was included in the acquisition time. After 3D acquisition was complete, clinicians restarted the timer and performed their standard 2D measurement workflow, consisting of locating and measuring the five largest follicles in each ovary together with the endometrial thickness.

We calculated the average 3D segmentation and analysis time for each substep in the 3D pipeline across 50 BICYCLE images.

Installation and code availability

All source code for FollicleFinder and for producing the plots in this manuscript is released under the 3-clause BSD license, and is publicly available as a git repository. One line installation and usage instructions are available on the README of each code repository.

- FollicleFinder (training and predicting 3D follicle segmentations, graphical user interface, publication figures): <https://git.bsse.ethz.ch/iber/follicle-finder>

- The deep learning weights of the fold 0 model that are trained during this study are openly available in the git repository under `src/follicle_finder/segmentation_model/l3d3dunet/weights/`

Results

Dataset

To help others further improve follicle segmentation algorithms, we are releasing our image dataset of raw images plus ground truth masks. The dataset contains 533 3D TVUS images annotated by medical doctors or ultrasound experts. These data were collected by vaginasonographic examination as part of the BICYCLE study²⁴. The release of this dataset represents a $15\times$ increase of 3D ultrasound images with segmentations from the previously published USOVA3D dataset²⁰. The image and ground truth segmentation data is available via the ETH research collection.

We also compared the 2D vs. 3D clinical timing and also evaluated the performance of GE Healthcare's Sono-AVC performance. Both of these analyses have been conducted under the human PROSPECT Study.

Quantitative benchmarking of ovarian follicle segmentation

To validate our segmentation algorithm, we measured performance using a test set consisting of 10% of our dataset and employed a 5-fold cross-validation strategy during training. To eliminate the risk of data leakage, all images from any individual patient were strictly assigned to either the same training fold or isolated entirely within the test set.

We evaluated object detection and segmentation performance using instance-level Precision and Recall on the test set. To capture instance-level clinical nuances, we classified error modes using the framework established by Greenwald et al.³⁵:

- True Positive (TP): A predicted object that shares an Intersection-over-Union (IoU) > 0.5 with a single matching ground truth object.
- False Positive (FP): A predicted object that does not match any ground truth object.
- False Negative (FN): A ground truth object that does not match any predicted object.
- Merge Error: A single predicted object encapsulates multiple ground truth objects by at least 50%.
- Split Error: A single ground truth object is broken up across multiple predicted objects by at least 50%.

To ensure statistical independence between traditional detection metrics and topological segmentation errors, instance-level Precision and Recall are calculated strictly using clean True Positive, False Positive, and False Negative counts:

- Precision = $TP / (TP + FP)$
- Recall = $TP / (TP + FN)$

Under this framework, instances flagged as Merge or Split errors are excluded from the True Positive count and are tracked entirely independently in Table 2. The reported merge and split rates represent the number of affected ground truth instances divided by the total number of ground truth instances.

We benchmarked FollicleFinder against a leading deeply supervised model, EXT 1+DS²¹, using the same BICYCLE dataset and their standard 80:20 train-validation split to ensure a direct, fair comparison (Tables 1 and 2). For broader context, we also include a secondary comparison with CR-Unet¹² (Table 1). However, these values are taken from the original publication and were obtained on a separate, proprietary dataset. Other recent

Table 1 Benchmarking of FollicleFinder performance versus the deeply supervised model “EXT 1+DS” from²¹ and CR-Unet from¹².

Model	Dataset	≥ 2 & < 5 mm		≥ 5 & ≤ 10 mm		> 10 mm	
		Precision	Recall	Precision	Recall	Precision	Recall
FollicleFinder	BICYCLE	0.9 ± 0.01	0.88 ± 0.0	0.97 ± 0.0	0.99 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
EXT 1+DS	BICYCLE	0.78	0.78	0.93	0.92	0.94	0.9
CR-Unet	Proprietary	0.96*	0.85*	0.99*	0.85*	0.88*	0.95*

Values are represented as mean $\pm$ standard deviation across all trained folds. (*) CR-Unet benchmarks are taken from their own manuscript and therefore not evaluated on the same test set.

Significant values are in bold.

Table 2 Merge and split benchmarking of FollicleFinder performance versus the deeply supervised model “EXT 1+DS” from²¹.

Model	Dataset	≥ 2 & < 5 mm		≥ 5 & ≤ 10 mm		> 10 mm	
		Merge	Split	Merge	Split	Merge	Split
FollicleFinder	BICYCLE	$4 \pm 1\%$	$1 \pm 0\%$	$8 \pm 2\%$	$4 \pm 1\%$	$1 \pm 3\%$	$4 \pm 1\%$
EXT 1+DS	BICYCLE	17%	0%	23%	2%	22%	0%

Values are percentages and if multiple folds were trained represented as mean $\pm$ standard deviation across all trained folds.

Significant values are in bold.

architectures, such as S-Net³⁶, were excluded from quantitative benchmarking because their authors did not report precision metrics or share their source code.

FollicleFinder consistently outperforms the deeply supervised EXT 1+DS model across all size cohorts at the instance level (Table 1), achieving near-perfect metrics in larger follicles. For small follicles, FollicleFinder achieves a 15% higher instance precision and a 13% higher recall over the EXT 1+DS model. For medium-sized follicles FollicleFinder exhibits excellent performance with a 4% higher precision and an 8% higher recall. Lastly for large follicles (> 10 mm) we achieve a flawless 1.0 in both precision and recall, notably outperforming EXT 1+DS with a 6% higher precision and 11% higher recall.

For merge errors, FollicleFinder shows a substantially lower merge rate than EXT 1+DS across all size categories, improving of 13, 15 and 21 percentage points for the small, medium and large size categories, respectively. Split errors remain low for both methods, averaging 3% for FollicleFinder and 1% for EXT 1+DS.

Although a direct comparison with CR-Unet is limited by differences in test sets and the unspecified event-calling methodology in their original publication¹², FollicleFinder achieves highly competitive or superior results despite being trained on only $\sim 19\%$ of the imaging volume used by CR-Unet (483 vs. 2509 images).

Overall, these results demonstrate that FollicleFinder delivers highly accurate instance segmentations, achieving the best performance on our evaluation dataset and placing it among the top-performing methods reported for this task. As explored in the following section, these minor error rates do not compromise FollicleFinder’s capacity to provide reliable clinical measurements, such as automated follicle counts and volumetric sizing.

Inter-operator segmentation agreement

Three ultrasound experts independently segmented all follicles in six 3D ovary volumes using our napari annotation plugin, without access to model predictions or each other’s labels. Agreement was assessed at the whole-ovary level using voxel-wise Dice on the complete follicle masks, and using ICC(2,1) (two-way random effects, absolute agreement, single rater) for per-image and per-follicle measurements.

Across all six cases, mean pairwise follicle mask Dice was 0.69 with standard deviation at 0.11, with an ICC(2,1) of 0.8 for follicle count and 0.91 for total follicle volume across cases. For the 30 follicle instances annotated by all three observers with one-to-one spatial correspondence, per-instance ICC(2,1) was 0.87 for

volume and 0.91 for effective diameter, showing that agreement remained high when comparing matched individual follicles rather than only case-level totals.

Agreement was lowest for expert 1 (mean pairwise Dice 0.68 vs. expert 2 and 0.66 vs. expert 3), who was segmenting 3D follicles for the first time and, on qualitative review, missed more follicles and produced more merge errors, particularly in lower-contrast volumes. This variance highlights how different operators, especially new operators will measure follicles differently which leads to suboptimal treatment outcomes due to standard rigid protocols relying on the measured follicle sizes.

As a comparison, all fold models across all test images have an average Dice score of 0.91 with a standard deviation of 0.05, which shows the robust performance of our model. This does not mean our model exceeds expert judgment on follicle presence. The difference in performance values should be interpreted cautiously, because manual 3D annotation is performed slice-by-slice using interpolated contours and a spherical brush, yielding irregular boundaries that reduce Dice relative to fully 3D algorithmic segmentations even when follicle identity is agreed. The inter operator Dice and the ICC(2,1) image level scores are therefore sensitive to boundary smoothness and interpolation artefacts, not only to disagreement about which follicles are present. As a conclusion, these results show a high variability between human operators and our high consistent model Dice scores reflect the robust 3D surface definition.

Clinically-relevant measurements

To characterize the performance of FollicleFinder on clinically-relevant metrics, we compared the values obtained from our fold 0 model predictions to the ground truth values. In particular, we compared the follicle counts (Fig. 4), follicle sizes (Fig. 5), and follicle localization (Fig. 6).

Follicle count and follicle sizes are measurements that can be used to inform in vitro fertilization treatment. For example, the number of antral follicles (2–5 mm in diameter) is a marker of the ovarian reserve and can be used as a predictor of the response to the hormonal stimulation used in IVF^{37,38}. Further, follicle size, is often used to guide the timing of the treatment in the controlled ovarian hyperstimulation step common in IVF³⁹. Thus, accurate measurements of follicle sizes and counts is important in clinical practice.

First, we compared the number of detected follicles to the number of follicles in the ground truth. FollicleFinder has a mean absolute error of 3.3 follicles (Fig. 4B), which is 50% higher than CR-Semi⁴⁰ (2.2 mean follicle error). FollicleFinder counts are well correlated with the ground truth and tend to perform less well in images with many small follicles.

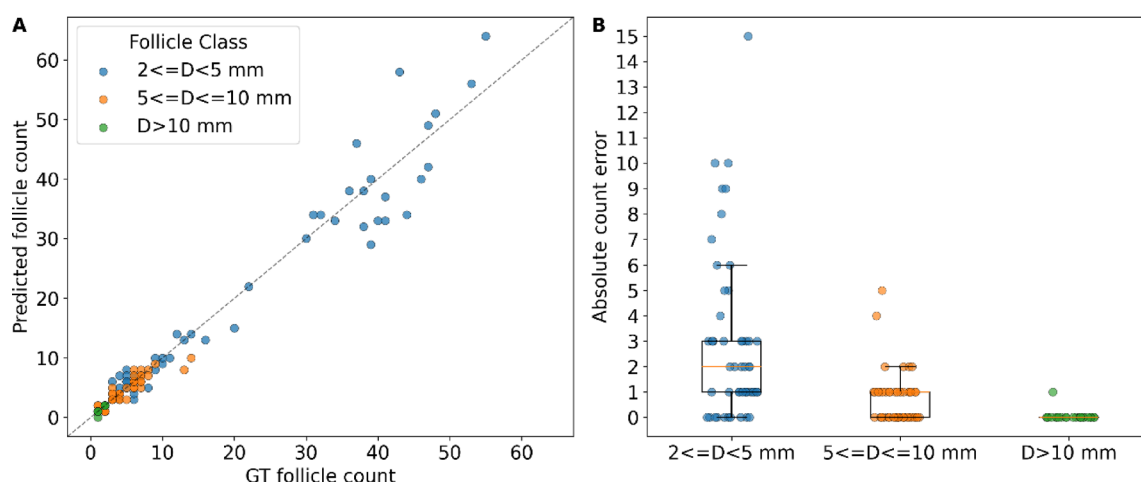


Fig. 4 (A) Predicted follicle counts are well-correlated with ground truth for all follicle classes. (B) Mean absolute count error is 3.3 follicles across all follicle classes. In the legend, D represents the effective diameter of each follicle.

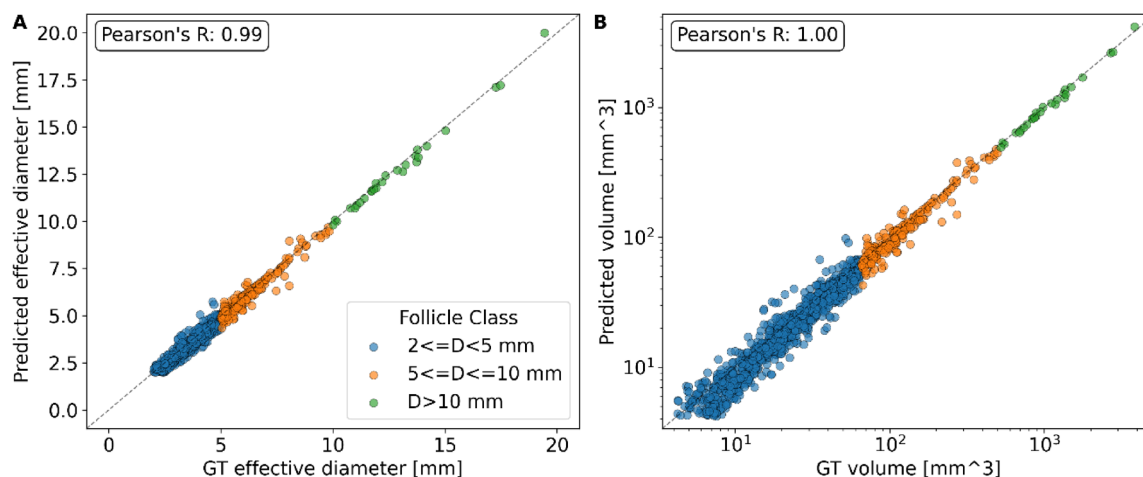


Fig. 5 (A) Measured effective diameters from the predictions are well-correlated with ground truth measurements. (B) Measured follicle volume from the predictions are well-correlated with ground truth measurements. Horizontal and vertical axes for the volume subplot are log scale. The gray dotted line in each plot is a line with a slope of 1 and an intercept of 0. In the legend, D represents the effective diameter of each follicle.

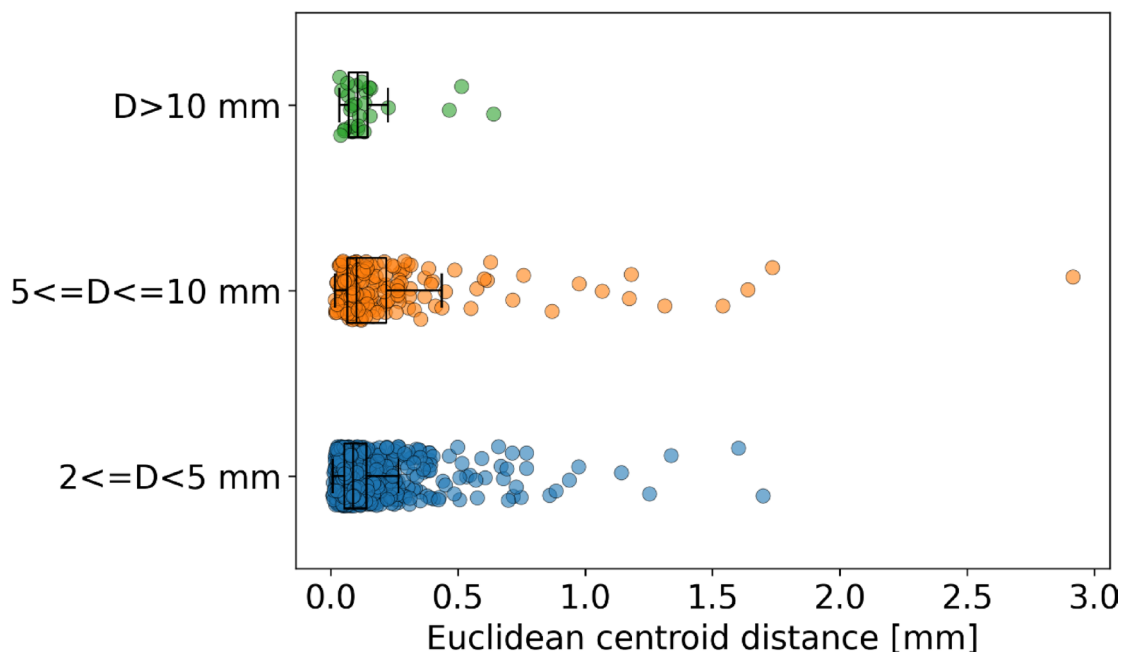


Fig. 6 Median follicle localization error is 0.09 mm, which is less than one voxel. In the legend, D represents the effective diameter of each follicle.

Next, we considered how accurately the size and location of the follicles was measured. To do so, we compared the effective diameter and volumes measured on the FollicleFinder predictions to those measured in the ground truth. For both effective diameter and volume, the measurements were highly correlated (Pearson's correlation coefficient, $r=0.99$, Fig. 5). Finally, we calculated the localization error when comparing the centroid detected by FollicleFinder to the centroid calculated from the ground truth. FollicleFinder had a median localization error of 0.09 mm, which is within one voxel (Fig. 6).

Additionally, we compared the measured follicle sizes to the manual measurements made by clinicians when the images were acquired. A standard clinical practice is to manually measure the biggest five follicles of each

ovary, judged by eye of the clinicians. The manual measurement of each follicle is made by averaging the major and minor axis of the follicle in the 2D plane where it appears biggest. Our dataset only contains the measurement of the biggest follicle. The equivalent 3D measurement is the average of the oriented bounding box extents of the segmented follicle. This measurement should coincide with the manual measurements if the clinicians find the optimal ultrasound probe orientation with the maximum 2D follicle cross section and then accurately measures the largest perpendicular follicle diameters. As the follicle shape becomes less spherical, the harder it becomes for clinicians to find the optimal ultrasound probe orientation. We therefore colored each follicle depending on their sphericity (1 corresponding to a perfect sphere), measured using Eq. (6).

$$sphericity = \frac{\pi^{1/3} * (6 * Volume)^{2/3}}{SurfaceArea} \quad (6)$$

When fitting a line to the comparison of the automated segmentation to the manual measurements, we obtained a slope of 1.09, which indicates that the manual measurements are overestimating the diameter of the follicles (Fig. 7). Further the Pearson's correlation coefficient is 0.96, which is lower than that when comparing the automated segmentations to the ground truth, indicating our measurements are better correlated with the ground truth measurements than the manual measurements. We additionally compared the manual measurements to our automated measurements and the ground truth over one cycle of a patient (Fig. 8). In agreement with the comparison

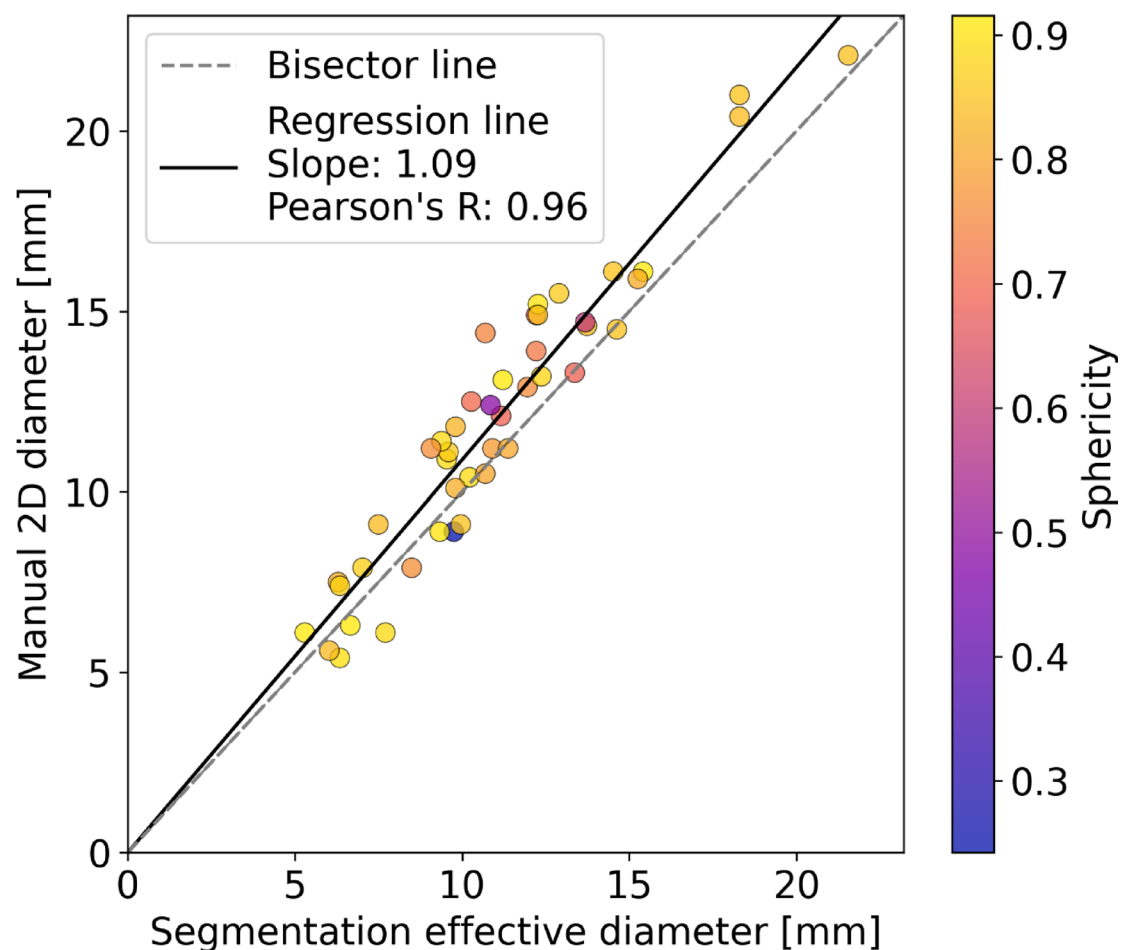


Fig. 7 Comparison to manual follicle measurements. The linear fit has a slope of 1.09 shows that the manual measurements tend to overestimate the follicle diameters.

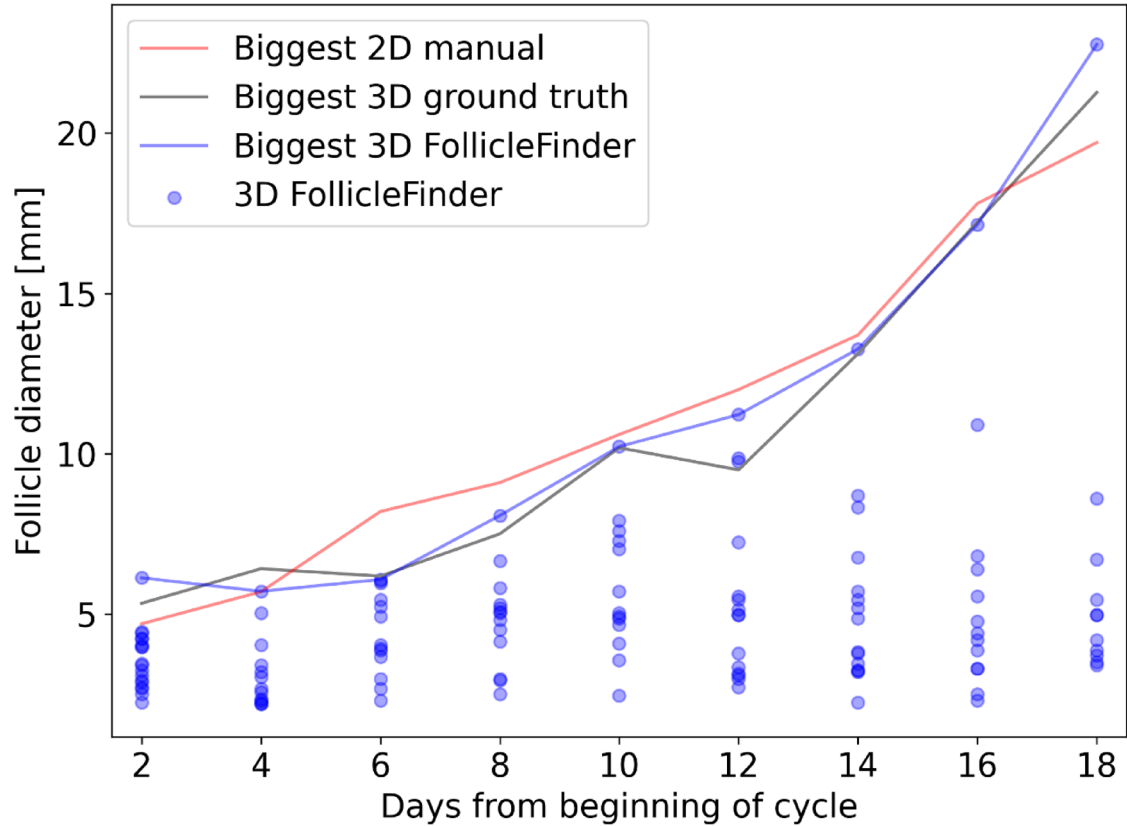


Fig. 8 Comparison of manual and automated diameter measurements over a cycle. Automated measurements enable profiling of all of the follicles. The manual diameter measurements were made a clinician by averaging two manual diameter measurements made in the plane in which the follicles appeared largest. The FollicleFinder measurements are the average of the dimensions of a bounding box fit to the automated follicle segmentation. The lines represent the biggest follicle at each time point and each point is an individual follicle.

Table 3 Comparison of acquisition and analysis times (mean and standard deviation in seconds) between the standard 2D workflow and the proposed 3D pipeline.

Steps	Mean	Std
2D Acquisition and Analysis	169	52
3D Acquisition	121	47
Pre-processing	4	1
LSD prediction	17	6
Segment follicles	31	12
Post-processing	17	7
Measurement	1	0
Total Single 3D Analysis	71	15
Total 3D Acquisition and Analysis	253	28

Significant values are in bold.

of manual and automated measurements in Fig. 7, the automated measurements follow the ground truth more closely than the manual measurements.

We evaluated the clinical efficiency of the proposed 3D ultrasound workflow against the conventional 2D standard of care (Table 3). The manual 2D workflow, encompassing real-time localization, slice selection, and measurement of the five largest follicles per ovary alongside the endometrial thickness, required a mean total time of 169 ± 52 s per examination. In contrast, the 3D volume acquisition phase alone (capturing the uterus, left ovary, and right ovary) required 121 ± 47 s, representing a 28.4% reduction in direct ultrasound examination

time. This acquisition time natively accounts for instances where clinicians required multiple attempts to secure high-quality diagnostic volumes.

The automated 3D analysis pipeline required a total mean processing time of 71 ± 15 s per volume, distributed across pre-processing, LSD prediction, follicle segmentation, post-processing, and automated measurement. When integrated directly into the ultrasound server, the system runs asynchronously to complete the dual ovarian analyses within 182 s and when including the uterine analysis within 253 s from exam initiation. Consequently, this autonomous workflow utilizes routine post-examination downtime, such as probe sanitization and patient dressing, to deliver fully synthesized diagnostic reports without prolonging the clinical encounter. We note that the times reported here describe image acquisition and downstream processing rather than the total time until diagnostic information reaches the clinician. The latter also depends on the local infrastructure, such as hardware, network bandwidth, and the reporting workflow, and may therefore vary between clinics.

FollicleFinder accurately extracts clinically relevant parameters from 3D ultrasound images. When comparing to both 3D ground truth and 2D manual measurements, FollicleFinder is able to accurately count and measure individual follicles, without negatively impacting imaging time. Future work will explore the accuracy and utility of other 3D morphometric features such as surface area in improving mechanistic models of folliculogenesis.

GE healthcare SonoAVC comparison

To contextualize our model's performance against existing commercial solutions, a preliminary qualitative evaluation was conducted using GE Healthcare's SonoAVC software¹³. Due to software compatibility and data-transfer limitations with the newer ultrasound hardware hosting SonoAVC, this evaluation was restricted to a representative subset of two high-quality 3D ovarian images independent of the main BICYCLE dataset. The commercial software exhibited notable segmentation errors; for instance, in one image, six of the ten largest identified 'follicles' were false positives located entirely outside the ovarian boundary with an effective diameter of up to 21.3 mm. Across both images a total of 39% (11 out of 28) of follicles were false positive objects that were outside the actual ovary and evenly distributed among the three size categories. In contrast, our proposed model demonstrated high anatomical fidelity, yielding 0% false positive rate for follicles > 10 mm outside the ground truth ovary, and 1.4% rate for 5–10 mm follicles and 3.0% for the 2–5 mm range. Additionally, SonoAVC failed to recognize and segment some follicle within the ovary, and in one case segmented a single follicle as two separate objects. Given the two-image sample, these results characterize SonoAVC's failure modes rather than its comparative accuracy; a matched benchmark was precluded by the constraints noted above.

Graphical user interface

To facilitate testing of segmentation algorithms and modeling integration in the clinic, we have also developed and released a GUI (Fig. 9). The FollicleFinder GUI wraps the segmentation and measurement pipeline in a napari plugin²⁶ such that it can be operated via a graphical user interface without requiring coding experience. Via the FollicleFinder GUI, users can load raw DICOM ultrasound images, run the segmentation and measurement pipeline, and interactively view the results. The user can explore the raw ultrasound image and segmentations in 2D and 3D, as well as view plots of the morphological measurements. By virtue of being implemented as a napari plugin, any python compatible computational folliculogenesis model can be easily integrated in FollicleFinder. In future work, we will explore directly integrating mechanistic models of folliculogenesis into FollicleFinder as well as other useful segmentation tasks such as segmenting the endometrium.

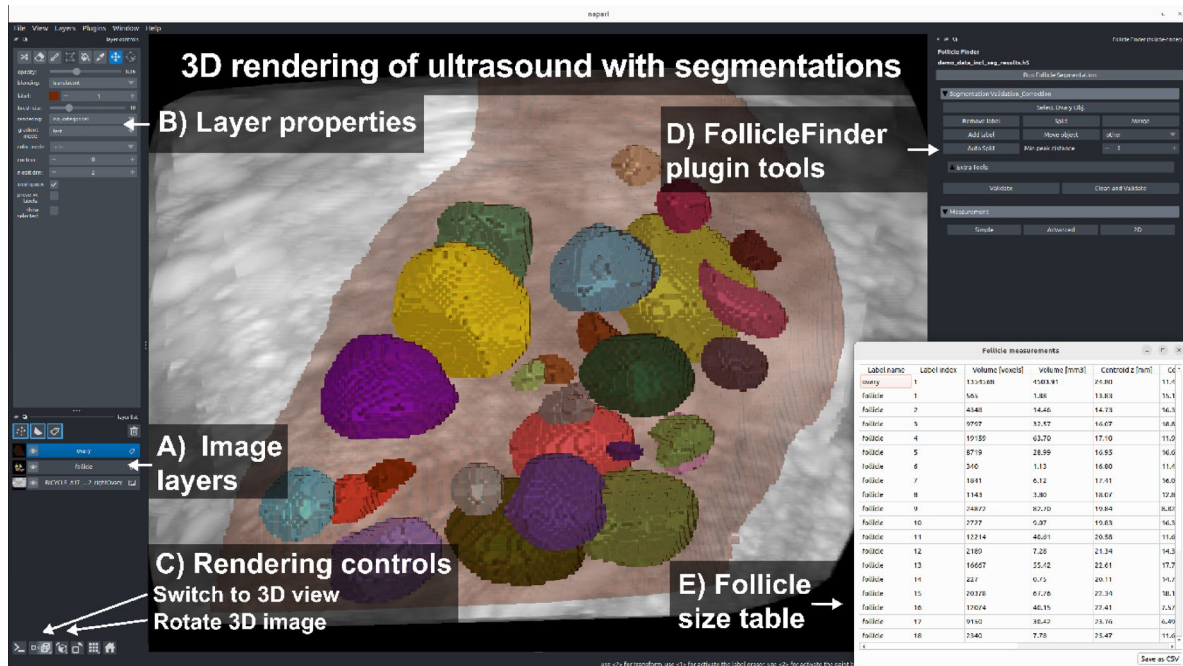


Fig. 9 The FollicleFinder napari plugin provides a graphical user interface for follicle segmentation. (A) Image layers: The three layers are the ovary mask, the follicle labels and the raw ultrasound image. (B) Rendering controls: There are rendering controls in the lower left-hand corner of the viewer window. (C) Layer properties: In the top left you can modify layer properties such as the layer opacity, which is useful for superimposing multiple layers and to only show the contour of labels. (D) FollicleFinder plugin tools: The panel on the right contains the run segmentation button, segmentation correction tools, and measurement tools. (E) Follicle size table: Following segmentation, you can run the measurement analysis and explore the results using the table that pops up as a separate window. In the current view you can see a 3D view of all the follicle instance segmentations over the respective ultrasound image. Note that in this view the ultrasound image cannot be interpreted.

Limitations

This study has several limitations. All BICYCLE images were acquired within a single clinic, on a single ultrasound machine and with a single probe type. How well FollicleFinder transfers to data from other probes, machines, and clinics therefore remains to be established, and some degree of retraining or recalibration is likely to be required, given the well-known problem of domain shift in medical image analysis⁴¹; however, our open-source code enables straight-forward retraining on new data. How much additional data from a new setting would be needed to reach comparable performance will be quantified in future multi-center work. We have not yet trained or evaluated our model on 3D ultrasound images from stimulated women with multiple large follicles in the same ovary. While we compared our results to manual measurements by clinicians, we did not assess the impact on clinical outcomes, as our focus here was on developing open-source pipelines that can be integrated into and built upon for personalized healthcare research. Finally, recent advances in image analysis, including image transformers⁴² and self-supervised pretraining methods⁴³, represent promising directions for future integration and evaluation.

Conclusion

In conclusion, we have developed a segmentation algorithm and accompanying software that achieves top-tier performance for clinically relevant measurements. We expect that model performance can be further improved by increasing the amount of training data. By releasing both our code, model and dataset as open resources, we

anticipate that other researchers will be able to build upon and improve these segmentation algorithms and to integrate the resulting rich 3D information into personalized health models and clinical practice.

Acknowledgements We thank Ting-Yu Ho and Steve Runser for helpful discussions, Dr. Ursula Gobrecht-Keller, Dr. Eva Maria Baumgartner-Sift, Dr. Giuseppina De Napoli, Dr. Bettina Dühler, Dr. James Geiger, Dr. Philipp Quaas from the University Hospital Basel for their help with the acquisition time analysis and Roman Vetter for proofreading the manuscript and providing helpful feedback.

Author contributions S.F., L.F., M.B., H.G. curated and manually annotated the ultrasound images; H.G. implemented the first segmentation methods; L.F., K.Y. implemented and trained the final pipeline, analysed and plotted the segmentation results. L.F. developed the graphical user interface. K.Y., L.F., D.I., C.D.G wrote the manuscript. All authors provided final approval of the manuscript for submission.

Funding This project was supported by the grant #2021–357 of the Strategic Focus Area “Personalized Health and Related Technologies (PHRT)” of the ETH Domain (Swiss Federal Institutes of Technology) and the European Union's Horizon 2020 research and innovation programme, Pan3DP FET Open [800981]. Data collection was funded by University Hospital Basel, Basel, Switzerland and Roche Diagnostics International Ltd, Rotkreuz, Switzerland.

Data availability All images and corresponding ground truth annotations from the BICYCLE Study that support the findings of this study, as well as the acquisition time data, and the SonoAVC results will be openly available via the ETH research collection with a doi upon publication of this manuscript. For the referees, a sample 3D ultrasound image file is available at https://git.bsse.ethz.ch/iber/follicle-finder/-/raw/main/src/follicle_finder/sample_data/3D_raw_ultrasound_image.dcm?ref_type=heads&inline=false.

Declarations

Competing interests All authors declare no competing interests.

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

References

1. Domar, A. D. Impact of psychological factors on dropout rates in insured infertility patients. *Fertil. Steril.* **81**, 271–273 (2004).
2. Pasch, L. A. et al. Addressing the needs of fertility treatment patients and their partners: Are they informed of and do they receive mental health services?. *Fertil. Steril.* **106**, 209–215 (2016).
3. Bachler, M., Menshykau, D., de Geyter, C. & Iber, D. Species-specific differences in follicular antral sizes result from diffusion-based limitations on the thickness of the granulosa cell layer. *Mol. Hum. Reprod.* **20**, 208–221 (2014).
4. Fischer-Holzhausen, S. & Röblitz, S. Hormonal regulation of ovarian follicle growth in humans: Model-based exploration of cycle variability and parameter sensitivities. *J. Theor. Biol.* <https://doi.org/10.1016/j.jtbi.2022.111150> (2022).
5. Iber, D. & de Geyter, C. Computational modelling of bovine ovarian follicle development. *BMC Syst. Biol.* **7**, 1–24 (2013).
6. Nisal, A., Diwekar, U. & Bhalerao, V. Personalized medicine for *in vitro* fertilization procedure using modeling and optimal control. *J. Theor. Biol.* **487**, 110105 (2020).
7. Shilo, M., Mayo, A. & Alon, U. A mechanism for ovulation number control. *Front. Endocrinol. (Lausanne)* **13**, 816967 (2022).
8. Fischer, S. et al. Mathematical modeling and simulation provides evidence for new strategies of ovarian stimulation. *Front. Endocrinol. (Lausanne)* **12**, 613048 (2021).
9. Mancini, T. et al. Computing personalised treatments through in silico clinical trials. A case study on downregulation in assisted reproduction. In *Intelligenza Artificiale* 1–16 (2018).

10. Lange, A., Schwieger, R., Plöntzke, J., Schäfer, S. & Röblitz, S. Follicular competition in cows: The selection of dominant follicles as a synergistic effect. *J. Math. Biol.* **78**, 579–606 (2019).
11. Röblitz, S. et al. A mathematical model of the human menstrual cycle for the administration of GnRH analogues. *J. Theor. Biol.* **321**, 8–27 (2013).
12. Li, H. et al. CR-Unet: A composite network for ovary and follicle segmentation in ultrasound images. *IEEE J. Biomed. Health Inform.* **24**, 974–983 (2020).
13. Raine-Fenning, N. et al. SonoAVC: A novel method of automatic volume calculation. *Ultrasound Obstet. Gynecol.* **31**, 691–696 (2008).
14. Noor, N., Vignarajan, C. P., Malhotra, N. & Vanamail, P. Three-dimensional automated volume calculation (sonography-based automated volume count) versus two-dimensional manual ultrasonography for follicular tracking and oocyte retrieval in women undergoing *in vitro* fertilization-embryo transfer: A randomized controlled trial. *J. Hum. Reprod. Sci.* **13**, 296–302 (2020).
15. Coelho Neto, M. A. et al. Counting ovarian antral follicles by ultrasound: A practical guide. *Ultrasound Obstet. Gynecol.* **51**, 10–20 (2018).
16. Liang, X. et al. CR-Unet-based ultrasonic follicle monitoring to reduce diameter variability and generate area automatically as a novel biomarker for follicular maturity. *Ultrasound Med. Biol.* **46**, 3125–3134 (2020).
17. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: A generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).
18. Hollandi, R. et al. nucleAIzer: A parameter-free deep learning framework for nucleus segmentation using image style transfer. *Cell Syst.* **10**, 453–458 (2020).
19. Mathur, P., Kakwani, K., Kudavelly, S., Ramaraju, G. A. & others. Deep Learning based quantification of ovary and follicles using 3D Transvaginal Ultrasound in Assisted Reproduction. in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* 2109–2112 (2020).
20. Potočník, B. et al. Public database for validation of follicle detection algorithms on 3D ultrasound images of ovaries. *Comput. Methods Programs Biomed.* <https://doi.org/10.1016/j.cmpb.2020.105621> (2020).
21. Potočník, B. & Šavc, M. Deeply-supervised 3D convolutional neural networks for automated ovary and follicle detection from ultrasound volumes. *Appl. Sci.* <https://doi.org/10.3390/app12031246> (2022).
22. Liu, G. et al. A weakly-supervised follicle segmentation method in ultrasound images. *Sci. Rep.* **15**, 13771 (2025).
23. Sheridan, A. et al. Local shape descriptors for neuron segmentation. *Nat. Methods* **20**, 295–303 (2022).
24. Biniasch, M., Laubender, R. P., Hund, M., Buck, K. & De Geyter, C. Intra-and inter-cycle variability of anti-Müllerian hormone (AMH) levels in healthy women during non-consecutive menstrual cycles: The BICYCLE study. *Clin. Chem. Lab. Med.* **60**, 597–605 (2022).
25. Fedorov, A. et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn. Reson. Imaging* **30**, 1323–1341 (2012).
26. Sofroniew, N. et al. napari/napari: 0.4.12rc2. Preprint at <https://doi.org/10.5281/zenodo.3555620> (2021).
27. Lee, K., Zung, J., Li, P., Jain, V. & Seung, H. S. Superhuman accuracy on the SNEMI3D connectomics challenge. *arXiv preprint arXiv:1706.00120* (2017).
28. Wolny, A. et al. Accurate and versatile 3D segmentation of plant tissues at cellular resolution. *Elife* **9**, e57613 (2020).
29. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 424–432 (2016).
30. Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Jorge Cardoso, M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* 240–248 (Springer, 2017).
31. van der Walt, S. et al. Scikit-image: Image processing in Python. *PeerJ* **2**, e453 (2014).
32. Attene, M. A lightweight approach to repairing digitized polygon meshes. *Vis. Comput.* **26**, 1393–1406 (2010).
33. Dawson-Haggerty et al. trimesh. Preprint at <https://trimsh.org/> (2019).
34. Looney, P., Stevenson, G. & Collins, S. 3D ultrasound file reading and coordinate transformations. *J. Open Source Softw.* **4**, 1063 (2019).
35. Greenwald, N. F. et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotechnol.* 1–11 (2021).
36. Mathur, P., Kakwani, K., Diplav, Kudavelly, S. & Ramaraju, G. A. Deep learning based quantification of ovary and follicles using 3D transvaginal ultrasound in assisted reproduction. In *42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society: Enabling Innovative Technologies for Global Healthcare Embc'20* 2109–2112 (2020).
37. Jayaprakasan, K. et al. Prediction of *in vitro* fertilization outcome at different antral follicle count thresholds in a prospective cohort of 1,012 women. *Fertil. Steril.* **98**, 657–663 (2012).
38. Bancsi, L. F. et al. Predictors of poor ovarian response in *in vitro* fertilization: A prospective study comparing basal markers of ovarian reserve. *Fertil. Steril.* **77**, 328–336 (2002).

39. Shrestha, D., La, X. & Feng, H. L. Comparison of different stimulation protocols used in in vitro fertilization: A review. *Ann. Transl. Med.* **3**, 137 (2015).
40. Yang, X. et al. Contrastive rendering with semi-supervised learning for ovary and follicle segmentation from 3D ultrasound. *Med. Image Anal.* **73**, 102134 (2021).
41. Guan, H. & Liu, M. Domain adaptation for medical image analysis: A survey. *IEEE Trans. Biomed. Eng.* **69**, 1173–1185 (2021).
42. Dosovitskiy, A. et al. An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* <https://github.com/> (2021).
43. Tang, Y. et al. Self-supervised pre-training of swin transformers for 3D medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 20730–20740 (2022).

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

Authors and Affiliations

Leopold Franz^{1,2} · Kevin A. Yamauchi^{1,2} · Selina Fricke^{1,3} · Marieke Biniasch³ ·
Harold Gómez^{1,2} · Christian De Geyter³ · Dagmar Iber^{1,2}

✉ Dagmar Iber
iberd@ethz.ch

¹ Department of Biosystems Science and Engineering, ETH Zürich, Schanzenstrasse 44, 4056 Basel, Switzerland

² Swiss Institute of Bioinformatics, Basel, Switzerland

³ Reproductive Medicine and Gynecological Endocrinology (RME), University Hospital, University of Basel, Basel, Switzerland