Custom Automatic Segmentation Models for Medicine and Biology based on FastSAM

SANTIAGO PARAMÉS-ESTÉVEZ[1,2], DIEGO PÉREZ-DONES[3,4], IGNACIO REGO-PÉREZ[5], NATIVIDAD OREIRO-VILLAR[5], FRANCISCO J. BLANCO[5], JAVIER ROCA PARDIÑAS[6], GERMÁN GONZÁLEZ PAZÓ[7], DAVID G. MÍGUEZ[3,4], ALBERTO P. MUÑUZURI[1,2]

[1]Group of NonLinear Physics,
University of Santiago de Compostela,
Facultade de Física, Rúa Xosé María Suárez Núñez, s/n, 15782, Santiago de Compostela,
SPAIN

[2]Centro de Investigación e Tecnoloxía Matemática de Galicia, CITMAga,
Plaza do Obradoiro, Colexio de San Xerome, s/n, 15705, Santiago de Compostela,
SPAIN

[3]Departamento de Física de la Materia Condensada,
University Autónoma de Madrid,
Avda. Reina Mercedes, s/n, 28792, Miraflores de la Sierra, Madrid,
SPAIN

[4]Centro de Biología Molecular Severo Ochoa,
University Autónoma de Madrid,
Calle de Lavoisier, 4, 28049, Madrid,
SPAIN

[5]Servicio de Reumatologia. GIR-INIBIC,
Hospital Universitario de A Coruña, Sergas, University of A Coruña,
As Xubias, 84, 15006 A Coruña,
SPAIN

[6]Department of Statistics and O.R. & SiDOR Group,
University of Vigo,
Faculty of Economic and Business Sciences, As Lagoas, Marcosende, 36310, Vigo,
SPAIN

[7]Healthcare innovation Advisor,
Merasys,
Avenida Ramiro Pascual S/N- Nave C 36213, Vigo, Pontevedra,
SPAIN

Abstract: - FastSAM, a public image segmentation model trained on everyday images, is used to achieve a customizable and state-of-the-art segmentation model minimizing the training in two completely different scenarios. In one example we consider macroscopic X-ray images of the knee area. In the second example, images were acquired by microscopy of the volumetric zebrafish embryo retina with a much smaller spatial scale. In both cases, we analyze the minimum set of images required to segmentate keeping the state-of-the-art standards. The effect of filters on the pictures and the specificities of considering a 3D volume for the retina images are also analyzed.

Key-Words: - Automatic segmentation, FastSAM, X-ray images, microscopy images, Low-Resource Friendly, Generalizable Approach.

Received: April 19, 2024. Revised: October 6, 2024. Accepted: November 9, 2024. Published: December 13, 2024.

1 Introduction

Training and designing an image segmentation model from scratch is not accessible to everyone or every project. In most cases, the problem is related to the impossibility of accessing large numbers of images. Achieving custom competent models requires experts, enormous amounts of data, computational power, and time. These resources are scarce for small companies and investigation groups, leading to an undesired imbalance in our globalized world.

It has been noted in the literature that, while large groups develop and furnish existing ideas with their abundant resources, underfunded small groups still are who, proportionally, propose more groundbreaking ideas that revolutionize science, as studied in [1]. One of the main advantages of large teams is the ability to access big databases and computational resources. Achieving similar results with less data and power would greatly benefit small groups and, therefore, science.

Automatic segmentation is becoming increasingly relevant, specifically in medical and biological environments, where large amounts of images must be processed to screen patients or to supervise the evolution of biological experiments, examples can be found in completely different fields ranging from cardiology, oncology, radiology, biology in general, cell segmentation, etc. as can be seen in studies [2], [3], [4], [5], [6]. Also, a study with more details on the benefits of implementing automatic segmentation in existing workflows can be found in [7].

In the absence of public or commercial tools for a specific case of study, by default, this task tends to be performed manually resulting, in some cases, in suboptimal health service or a reduction in the scope of the biological studies. Also, some useful data can be abandoned instead of being used to train models to accelerate or even completely automatize acquisition or labeling processes. With this work, we would like to help small groups develop their custom models by showing the viability of our approach for creating performant models with as few resources as possible.

In this direction, several studies have been conducted to apply recent general segmentation models like the Segment Anything Model (SAM) from MetaAI, described in [8]. In [9], SAM is directly evaluated on tens of thousands of medical images extracted from openly available datasets. The model requires the user to specify points or regions of interest to segment the desired object. This general approach works very well in normal photographs, where objects are usually easy to identify. Nevertheless, for medical or biological images, the result is not always as good as expected or requires too much user input to be used in an automated framework. The alternative is to finetune SAM and for that in [10] a dataset of 1.5 million images was developed and used in conjunction with 20 A100 GPUs to train and test MedSAM. This model is a finetuned version of SAM that achieves high performances at segmenting several kinds of medical conditions (tumors, cuts, dark spots, etc.) and image types (X-ray, CT, MRI, etc.). It should be noted that having access to all those resources is not trivial, even for big companies. Therefore, discovering approaches to achieve similar results with a small fraction of that computational power and data has a lot of interest for reproducibility and potential future studies. That is why we will explore in this study the possibilities of using FastSAM for similar purposes since it claims to be a lighter alternative to SAM as shown in [11].

This manuscript will use two completely different sets of images to prove this approach's viability. On the one hand, knee X-ray images from the OAI (OstheoArthitis Initiative), presented in [12] and on the other, microscopical images of the 3D retina nuclei from a zebrafish embryo. In the first case, the bones observed are, in most of the cases, clearly separated from the surrounding tissue while, in the second example, the objects to analyze are composed of a myriad of small objects, thus, complicating the task of recognizing the ensemble for a non-trained eye.

Our approach consists of finetuning FastSAM with unconventional data to see if its behavior can be generalized to other more scientific non-trivial settings. FastSAM can tell apart objects with clear boundaries, but more subtle cases, that require instruction even for human eyes, are more complicated to solve.

To tackle this problem with as little data as possible, we will trade FastSAM's generality for the performance at finding a single type of object, reducing drastically the resources needed to achieve significant results.

In this work, we demonstrate how FastSAM, a public image segmentation model trained on everyday images, can be used to achieve a customizable and state-of-the-art segmentation model with very few resources.

The next section describes the images used and their specificities. Also, the parameters that describe the goodness of the training are introduced in this section. The following section shows the results for the two cases considered and the manuscript finishes with a discussion and conclusions section.

2 Background and Methodology

To approach the problem of giving access to custom segmentation models to small groups, we tried to move away from conventional segmentation models such as U-Net, which are more challenging to implement and may require the presence in the research group of a person with experience in artificial intelligence, which may not be available to all small groups. Nevertheless, in [13], similar efficiencies for FastSAM and U-Net were observed when segmenting brain tissue (Dice score of ~0.95). Said work relied on the general capabilities of the pre-trained FastSAM model to segment the visible brain during surgery. This was done by giving it an ROI (region of interest), which in their setting is easier to set since the camera is always fixed. If the camera were to move, this process would benefit from finetuning FastSAM as we will show in this work.

Another alternative, as commented in the previous section, is using directly SAM, but this also has some disadvantages, like how heavy the model is to train and evaluate. Also, in [13], an implementation with SAM was made, but it was determined impractical due to high computational times, a lot greater than the time needed to do a manual segmentation. This shows the relevance of evaluation speeds for our purpose, the models obtained must be fast to be useful.

FastSAM is intended to be a model capable of segmenting any object with a prompt from the user, just like SAM, except FastSAM also allows text prompts. When fine-tuning FastSAM to only recognize one class of name “object”, the human interaction is removed by simply prompting said keyword to get the segmentation.

All of this makes FastSAM a very powerful candidate for custom design and automatic segmentation models with few resources as we will show in this section.

2.1 Model

FastSAM is a model based on YOLOv8, presented in [14], that has been trained with 2% of all SA-1B datasets. It achieves similar results to SAM, as shown in [8], but 50 times faster. To finetune it, the datasets must be crafted in the COCO format, described in [15].

The model is designed so it can take images with any aspect ratio. To do so, the shortest dimension is padded to create a square image and then scaled to the size specified when loading the model to train or evaluate. This is particularly useful for evaluating X-ray images from different equipment, which may have varying resolutions and shapes.

The models were trained in a node with 5 Nvidia A100 GPUs hosted by CESGA. All the models trained for this work were evaluated on a desktop computer with an Intel i5-13500 CPU (notice that this is common equipment for any lab). The parameters chosen to train the models and approximate training and inference times can be seen in Table 1.

Table 1. Summary of the parameters chosen to train models with each type of image. Image size has a noticeable impact on training time, but inference time per image is unchanged.

Model	Train Configuration			Times	
	Epochs	*Image Size*	*Batch*	*Train (min)*	*Evaluate (s/image)*
Tibia	200	1536x1536	30	~50	~5
Retina	200	1024x1024	40	~15	~5

2.2 Data Acquisition and Preprocessing

We will be working with large-scale X-ray images and microscopical images of a 3D object to show the viability of this method independently on the application, scale of the objects analyzed, etc.

X-ray data was acquired from the OAI, a database with more than 4000 patients with images periodically taken over up to 10 years with several devices at different hospitals. Each image has two knees, to train the model they were separated into two images, so the model only sees one knee at a time. Originally, images were stored in 16-bit but later converted to 8-bit RGB images in grayscale (same value for each pixel at the three channels).

Suboptimal acquisition conditions (i.e., wrong placement of the patient in the X-ray machine, non-standard functional settings of the machine, existence of spurious objects, etc.) can lead to the appearance of fog and illumination gradients in X-ray images, hiding from the human eye features that can be hard to see even under normal conditions. To aid in the manual segmentation of tibias, the knee images were enhanced with CLAHE (Contrast Limited Adaptive Histogram Equalization) which has been reported to be a good filter for medical

images in [16], [17]. This filter reduces both effects, clarifying most of the cases.

The segmented region of the tibia excludes areas that become visible when the knee is slightly rotated during the image acquisition. Under ideal conditions the front and the back of the tibia's head align, showing a clear border in the image. When the bone is rotated (due to imperfections during the image acquisition process or to the non-standard shape of the patient's knee) the desired border becomes a lot less visible due to the superposition. The images were meticulously segmented in collaboration with a medical specialist to find this border. The process was repeated for the femur, but its discussion will be omitted, since it must be segmented as a whole, and the problem is not as challenging (nevertheless a summary is presented in Appendix A).

A transgenic zebrafish line was used in the experiment to obtain the other set of images on the microscopic scale, expressing a fusion protein composed of histone2b and RFP which labels all nuclei in the fish. Embryos were collected at around 20 hours post fertilization, based on visual inspection, as specified in [18] and immobilized for imaging in a 1.5% agarose gel matrix. Images were taken in a confocal microscope STELLARIS 8 coupled to an inverted microscope model DMi8 (Leica), capturing the whole volume of the retina, with a section thickness of 1μm and overlapping between sections of 0.2μm. Typically, 151 retina sections are taken at every instant of time, the first corresponding to the upper part of the retina and then moving deeper into it. Frames were taken with a 1h time-step between them. A white laser diode of the microscope was adjusted to emit at 555 nm in wavelength to maximize fluorophore excitation and emission.

Histone 2b is a well-characterized protein produced by all cells which works as a scaffold for chromatin packaging, as described in [19]. This protein, along with other histones, has been extensively used along with fluorescent proteins, to label cell nuclei in in vivo tissues.

In our specific case, this histone has been fused with RFP, a fluorescent protein that has reddish fluorescence when excited, described in [20]. The general idea of the process is, as H2B protein is expressed in all cells and located in their nuclei, when the fusion protein is exposed to a specific light wavelength (555nm) the fluorophore emits fluorescence in a different wavelength (583nm) which is then captured by the microscope camera. By this means, each H2B-RFP present in the nuclei of the tissue would emit a fluorescent signal that can be later analyzed.

In other words, the fusion protein provides two different things, first H2B gives certainty of nuclear localization of the fluorescent protein, whereas RFP is the one in charge of providing a signal which can be captured in a fluorescence microscope.

The animals are maintained and bred according to protocols established in [21].

To improve cell nuclei visibility, images were also enhanced with a global histogram contrast equalization and then normalized. This effect is most noticeable at the bottom slices of each time frame, where light has to transverse more tissue and gets absorbed easily due to light scattering, letting less light reach the microscope.

2.3 Statistical and Control Parameters

To evaluate the performance of the trained models we have chosen as evaluation parameters precision (fraction of the prediction correctly guessed), recall (fraction of the ground truth correctly predicted), and the Dice or F1 score, as seen in the bibliography [2], defined in equations (1) and (2),

$$Precision = \frac{TP}{TP+FP};\ Recall = \frac{TP}{TP+FN}, \tag{1}$$

$$Dice\ score = \frac{2TP}{2TP+FP+FN}, \tag{2}$$

where TP, TN, FP, and FN are the fractions of pixels classified as true positives, true negatives, false positives, and false negatives, respectively [2], [3] .

The performances of the retina models were also compared with the mean image brightness and the number of retina nuclei in each slice.

The number of objects per slice was obtained using an in-house developed algorithm, described in [22] and based on top-hat transform. To count the retina nuclei, the algorithm was fed with the images intersected with the ground truth, so only the nuclei of interest can be counted.

2.4 Dataset Creation

All the segmentations were stored as binary images and converted to the COCO dataset format to train FastSAM, to generate masks with holes for the retina dataset, code from [23] was used.

A dataset consists of two folders, one with the images and the other with a text file per image, where the contour of a mask is specified as a row per object. The first number of each row represents the class associated with its object, while the rest are its contour points stored as alternating pairs of x and

y normalized coordinates (x_1,y_1,x_2,y_2,…,x_n,y_n). In this work training samples only have one object per image and models were trained with only one class. These changes greatly reduce the complexity of the task, reducing the amount of information required to build a performant model.

The total number of segmented images is 166 tibias and 1960 retinas. Since the objective is to achieve usable models with a training set of images as small as possible, only some of those images were used to train the models with increasing amounts of training samples.

The number of samples used to train each model was increased progressively by adding more images to the original set (3, 15, 41, 83, and 125 for the tibia and 124, 248, and 372 for the retina). The data that was not included in the training sets was used to test all the models, keeping the evaluation set constant ensures all the models are evaluated under the same conditions. FastSAM was finetuned with the original images as well as with their enhanced versions to study if the improvement in visibility for the human eye also helps FastSAM achieve better results.

The combination of all these changes results in a total of 16 different models, 10 for tibias and 6 for retinas, that will be evaluated in the following section.

3 Results

A total of 16 models were evaluated using their corresponding test datasets (10 for tibias and 6 for retinae). Dice score, recall, and precision values for each one can be found in Fig. 1, it seems the mean performance of all the tibia models is almost independent of whether it was filtered or in the number of images used to train the model. This is not the case for the retina model, which seems to benefit from using enhanced images or more training data. More information can be found in Table 2, the average of each one of the distributions shown in Fig. 1 is indicated, these parameters have also been calculated using femur segmentations as shown in the Appendix, in Table A1 and Figure A1 in Appendix, respectively. This gives a numerical reference of the performance for all the trained models. Increasing the number of training images has a slight positive effect on the performance of tibia models shown by a subtle increase in the average dice score. The effect of enhancing the tibia images is almost imperceptible, with 3 train samples the performance decreases compared with its non-enhanced counterpart, but for the rest of the models, the performance seems to be independent of the number of training samples.

A possible explanation for this phenomenon could be that standardizing the brightness reduces, even more, the complexity of the dataset, making the gradual progression observed for the non-enhanced case imperceptible. In other words, a performant model is achieved faster. This effect is also present in the retina models. The non-enhanced model with fewer train images has much worse performance than its enhanced counterpart, probably due to the diversity in brightness in the retina set of slices, dependent on depth and time of acquisition. This can be seen, for example, in Fig. 3a, where the image is brighter the closer it gets to the eye surface (slice 15) and darker at the deepest slices (slice 140). In contrast, all enhanced retina images (i.e. Fig. 3d) have similar brightness levels and therefore the models can generalize more easily.

Graphic comparisons between enhanced and non-enhanced models are shown in Fig. 2 for tibia and, in the Appendix, Figure A2, for femur. Once again knee model performances are almost invariant, while the retina models seem to benefit from increasing the number of patients. Once again, this makes sense, since one frame (the whole 3D image of the retina) has 151 different slices, training with fewer data means more extrapolation to unseen slices and thus, more failure. Also, the set of slices obtained for a retina are quite different depending on their position, therefore, a significant number of slices covering the whole retina needs to be used for training to achieve acceptable results. We observe that for all the cases considered the performance of the models is very high, only in the retina images the performance is observed to improve significantly as the number of training images increases up to 248 and ahead. Also, note that filtering the images to enhance the contrast did not result in any significant advantage (except for the one trained with 124 zebra fish images). To summarize, the performance of a retina model trained with few images is far worse than its analogous for tibias. This is related to the fact that each zebrafish image is taken at a different focal plane and, thus, effectively corresponds with a different object as the retina geometry differs greatly.

All knee models have very similar and high performances and this is exemplified in Fig. 2, where a slight improvement in the dice score can be appreciated with the increase in the number of patients used to train. This is particularly relevant in the areas close to the boundary, which are better detected as the number of training images increases.

Since the most difficult section of the tibias is in the head, to check for a bias in the scores due to long straight bone regions, they were recalculated for each model using only the upper third of each segmentation. Since the results were almost identical to those shown in Fig. 1, they have been omitted.

The retina model trained with the least amount of non-enhanced images had the lowest performance overall (Fig. 3a), where the predicted masks couldn't accurately find the retina nuclei. With more training images, the retina is properly recovered.

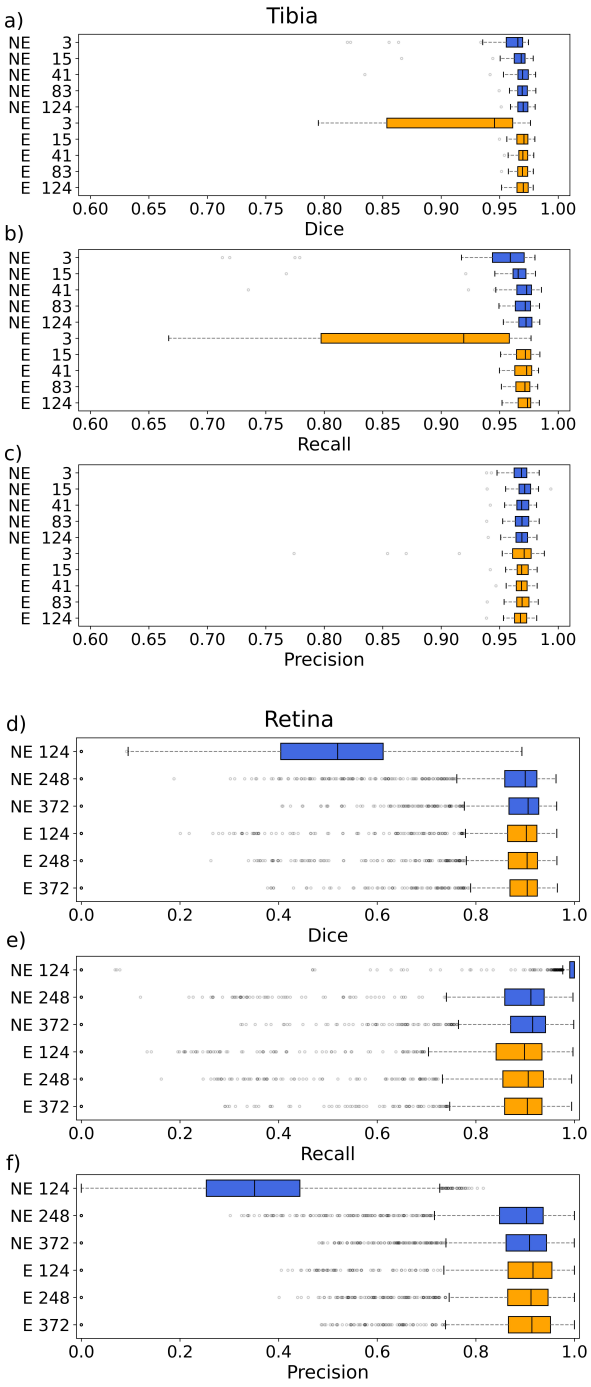


Fig. 1: Comparison of the performance of each model with enhanced and non-enhanced images. (a, b, c) Tibias. (d,e,f) Retinae. The distribution of values is shown with box plots, which mark the values from Q1 to Q3 with a box, the median is marked with a black vertical line in each box. Points found further than 1.5 times the box length are marked as outliers with small translucent circles. In blue, models trained with Non-Enhanced (NE) images, while in orange; with them Enhanced (E)

Table 2. Summary of the Dice, Recall (Rec.), and Precision (Prec.) average values achieved for each model

		Non Enhanced			Enhanced		
	Training images	***Dice***	***Rec.***	***Prec.***	***Dice***	***Rec.***	***Prec.***
Tibia	3	0.951	0.939	0.967	0.811	0.786	0.960
	15	0.965	0.960	0.971	0.969	0.970	0.969
	41	0.966	0.964	0.968	0.969	0.970	0.968
	83	0.969	0.970	0.968	0.969	0.969	0.968
	124	0.970	0.972	0.968	0.969	0.972	0.967
Retina	124	0.503	0.925	0.360	0.856	0.848	0.897
	248	0.843	0.856	0.872	0.870	0.868	0.886
	372	0.869	0.875	0.889	0.862	0.860	0.899

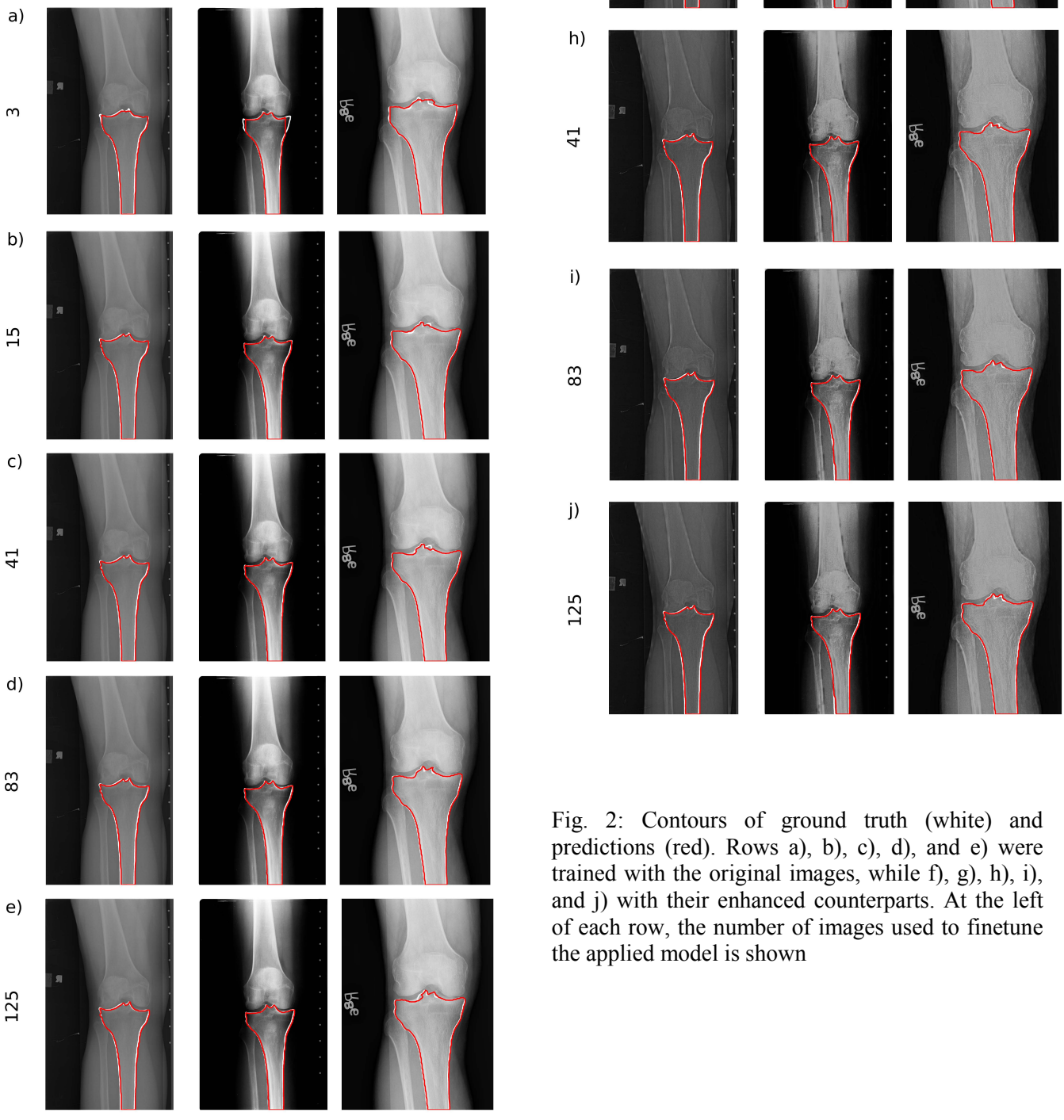


Fig. 2: Contours of ground truth (white) and predictions (red). Rows a), b), c), d), and e) were trained with the original images, while f), g), h), i), and j) with their enhanced counterparts. At the left of each row, the number of images used to finetune the applied model is shown

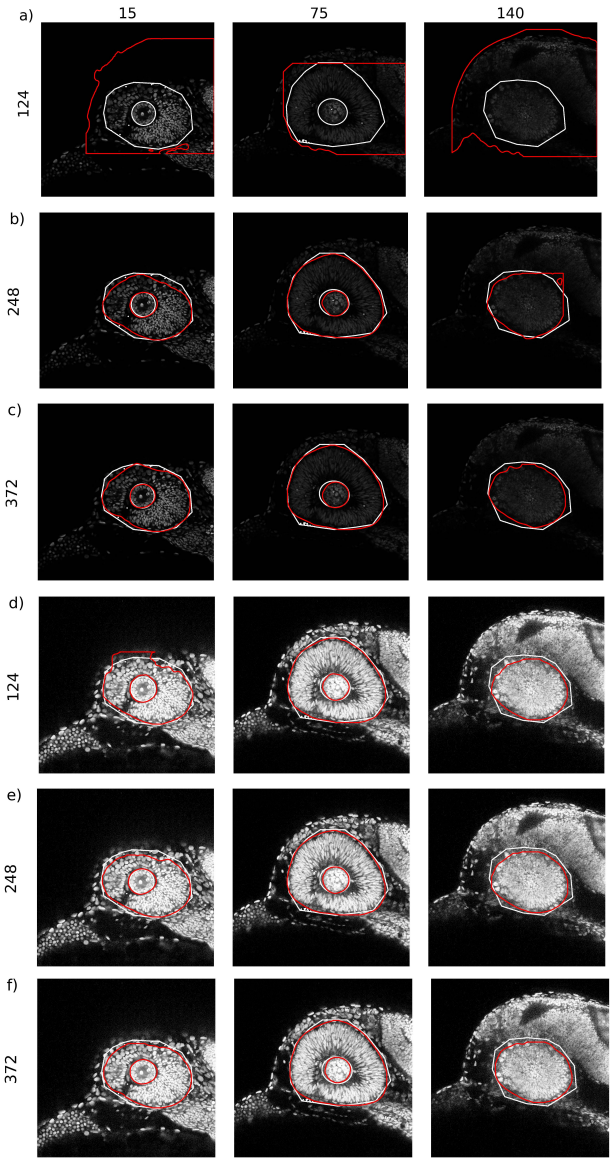


Fig. 3: Contours of ground truth (white) and predicted masks for each model (red). To illustrate the performance of the model at three sections of the retina (beginning, middle, and end), the index of the corresponding slices has been marked with its index on top of each column, the total number of slices per frame is 151. Rows a), b), and c) were evaluated with models trained and evaluated with 124, 248, and 372 of the original images, respectively, while rows d), e), and f); with 124, 248, and 372 of the enhanced images

It is important to note that, for this example, the retina is composed of many small-size objects (making it difficult for the non-trained eye to detect them) but the models can recover the structure with high precision when the number of training images is above 228.

As indicated in the methods section, the retinae correspond with a 3D structure, thus, for each experiment a stack of images is acquired corresponding with different sections of the retina. Typically, between 100 and 200 sections are acquired. In the present case, 151 slices were captured at each instant of time. Fig. 4 shows the quality indicators for each frame averaging over 11 different acquisitions comparing enhanced and non-enhanced images. For each slice (corresponding with a particular section of the retina) we plot the dice score, the average light intensity in the image, and the number of nuclei (the small objects that constitute the whole retina). Note that when non-enhanced images are used (Fig. 4a), the Dice score does not perform well at all depths, in particular, the first and latest slices are badly recovered. Nevertheless, when the initial enhancing filter is applied (Fig. 4b), the dice score improves, and the latest slices are now well recovered. This indicates that if the number of constituents of the retina (nuclei) is large (as it happens in the latest slices), the contrast enhancement may play a significant role, while for the first slices as the number of nuclei is small, increasing the contrast does not improve the result.

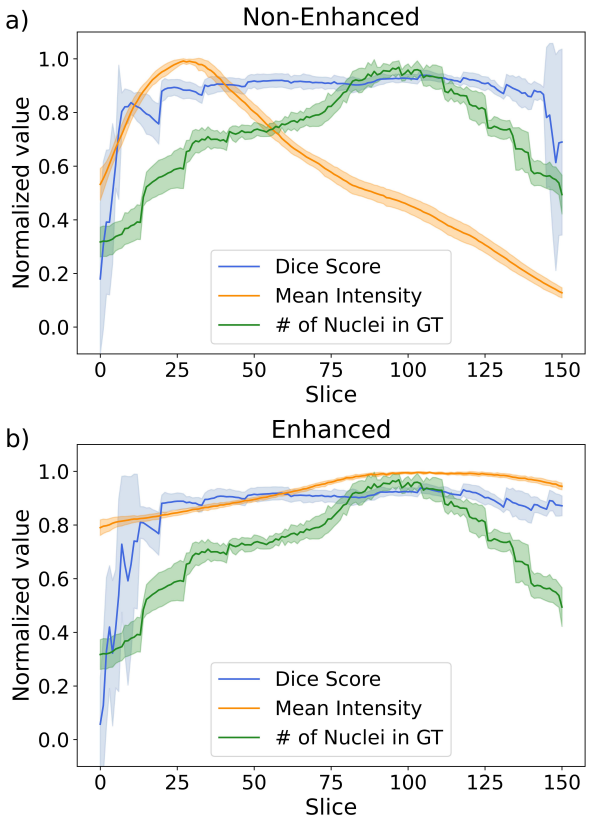


Fig. 4: Several retina image properties are compared with the Dice scores at each frame

The curves were obtained by averaging the values measured at each slice at 11 human-labeled

frames, the shaded regions correspond to the values closer to one standard deviation from the mean. Both models were trained with 372 images. Pixel intensity and number of cells were normalized. a) shows values obtained for non-enhanced images, while b) for their enhanced counterpart.

In this section, we have characterized the performance of finetuned models for each dataset and how the variation of some of the parameters, like the number of train samples or the preprocessing, affect their overall performance. We have also proven how our approach can be applied to both 2D and 3D images, very common types of data in both medical and biological settings, but also several branches of science and engineering. As a result, we have obtained very performant, specific, and light models that anyone with enough labeled images can reproduce.

4 Conclusion

We demonstrate in this manuscript that the use of a model trained with general images such as FastSAM, makes it possible to achieve competitive results for specific non-trivial applications. We consider two examples with images acquired by completely different means and scales, corresponding to situations with several difficulties. In both cases the protocol succeeded, and the segmentation was possible even after training the models with a small number of images. Additional cases such as the femur (also present in the X-ray images) were also analyzed and successfully segmented (see results in Appendix A).

With the zebrafish retina, we have also shown that for volumetric images a single model is enough to obtain a performant segmentation model. And that experimental image acquisition conditions can be filtered to improve the performance of the model.

This study also shows how finetuning FastSAM can be advantageous since it gives a lot of flexibility in the format of the input data, and also is faster and cheaper to achieve than finetuning SAM. It also requires less knowledge than training a U-Net (or finetuning if a compatible model is available) or other typical image segmentation models from the literature. All these factors prove once again that this approach is not only very viable for low-resources or non-expert groups but also useful to expand the frontiers of science, giving power to those who lack resources but exude innovation.

Following the approach presented in this contribution, it is possible to envision a path to develop equivalent tools for many diverse applications in a great variety of systems such as the ones presented here. We demonstrated that our training mechanism achieves results with equivalent accuracy to other non-pretrained methods, thus, competing with state-of-the-art models by taking advantage of all the information already embedded in FastSAM.

Acknowledgement:
Model training was done at CESGA (Center of Supercomputation of Galicia) and we acknowledge their support.

References:

[1] L. Wu, D. Wang, and J. A. Evans, “Large teams develop and small teams disrupt science and technology,” *Nature*, vol. 566, no. 7744, pp. 378–382, Feb. 2019, doi: 10.1038/s41586-019-0941-9.

[2] B. Serrano-Antón, A. Otero-Cacho, D. López-Otero, B. Díaz-Fernández, M. Bastos-Fernández, V. Pérez-Muñuzuri, J. R. González-Juanatey, and A. P. Muñuzuri, “Coronary Artery Segmentation Based on Transfer Learning and UNet Architecture on Computed Tomography Coronary Angiography Images,” *IEEE Access*, vol. 11, pp. 75484–75496, 2023, doi: 10.1109/ACCESS.2023.3293090.

[3] S. P. Primakov, A. Ibrahim, J. E. van Timmeren, G. Wu, S. A. Keek, M. Beuque, R. W. Y. Granzier, E. Lavrova, M. Scrivener, S. Sanduleanu, E. Kayan, I. Halilaj, A. Lenaers, J. Wu, R. Monshouwer, X. Geets, H. A. Gietema, L. E. L. Hendriks, O. Morin, *et al.*, “Automated detection and segmentation of non-small cell lung cancer computed tomography images,” *Nat Commun*, vol. 13, no. 1, p. 3423, Jun. 2022, doi: 10.1038/s41467-022-30841-3.

[4] P. Cheng, Y. Yang, H. Yu, and Y. He, “Automatic vertebrae localization and segmentation in CT with a two-stage Dense-U-Net,” *Sci Rep*, vol. 11, no. 1, p. 22156, Nov. 2021, doi: 10.1038/s41598-021-01296-1.

[5] T. Piotrowski, O. Rippel, A. Elanzew, B. Nießing, S. Stucken, S. Jung, N. König, S. Haupt, L. Stappert, O. Brüstle, R. Schmitt, and S. Jonas, “Deep-learning-based multi-class segmentation for automated, non-invasive routine assessment of human pluripotent stem cell culture status,” *Comput*

Biol Med, vol. 129, p. 104172, Feb. 2021, doi: 10.1016/j.compbiomed.2020.104172.

[6] C. Wen, M. Matsumoto, M. Sawada, K. Sawamoto, and K. D. Kimura, "Seg2Link: an efficient and versatile solution for semi-automatic cell segmentation in 3D image stacks," *Sci Rep*, vol. 13, no. 1, p. 7109, May 2023, doi: 10.1038/s41598-023-34232-6.

[7] G. R. Sarria, F. Kugel, F. Roehner, J. Layer, C. Dejonckheere, D. Scafa, M. Koeksal, C. Leitzen, and L. C. Schmeel, "Artificial Intelligence–Based Autosegmentation: Advantages in Delineation, Absorbed Dose-Distribution, and Logistics," *Adv Radiat Oncol*, vol. 9, no. 3, p. 101394, Mar. 2024, doi: 10.1016/j.adro.2023.101394.

[8] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment Anything," *ArXiv*, Apr. 2023.

[9] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment Anything Model for Medical Image Analysis: an Experimental Study," *ArXiv*, Apr. 2023, doi: 10.1016/j.media.2023.102918.

[10] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nat Commun*, vol. 15, no. 1, p. 654, Jan. 2024, doi: 10.1038/s41467-024-44824-z.

[11] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang, "Fast Segment Anything," *ArXiv*, Jun. 2023.

[12] G. Lester, "The Osteoarthritis Initiative: A NIH Public–Private Partnership," *HSS Journal*, vol. 8, no. 1, pp. 62–63, Feb. 2012, doi: 10.1007/s11420-011-9235-y.

[13] C. Li, X. Fan, R. B. Duke, K. L. Chen, L. T. Evans, and K. D. Paulsen, "Intraoperative stereovision cortical surface segmentation using fast segment anything model," in *Medical Imaging 2024: Image-Guided Procedures, Robotic Interventions, and Modeling*, SPIE, Mar. 2024, p. 21. doi: 10.1117/12.3006873.

[14] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," *GitHub*. Github, 2023. Accessed: May 29, 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[15] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," *ArXiv*, May 2014.

[16] R. Sharma and A. Kamra, "A Review on CLAHE Based Enhancement Techniques," in *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, IEEE, Sep. 2023, pp. 321–325. doi: 10.1109/IC3I59117.2023.10397722.

[17] B. Kurt, V. V. Nabiyev, and K. Turhan, "Medical images enhancement by using anisotropic filter and CLAHE," in *2012 International Symposium on Innovations in Intelligent Systems and Applications*, IEEE, Jul. 2012, pp. 1–4. doi: 10.1109/INISTA.2012.6246971.

[18] C. B. Kimmel, W. W. Ballard, S. R. Kimmel, B. Ullmann, and T. F. Schilling, "Stages of embryonic development of the zebrafish," *Developmental Dynamics*, vol. 203, no. 3, pp. 253–310, Jul. 1995, doi: 10.1002/aja.1002030302.

[19] S. P. Khare, F. Habib, R. Sharma, N. Gadewal, S. Gupta, and S. Galande, "HIstome--a relational knowledgebase of human histone proteins and histone modifying enzymes," *Nucleic Acids Res*, vol. 40, no. D1, pp. D337–D342, Jan. 2012, doi: 10.1093/nar/gkr1125.

[20] A. Miyawaki, D. M. Shcherbakova, and V. V Verkhusha, "Red fluorescent proteins: chromophore formation and cellular applications," *Curr Opin Struct Biol*, vol. 22, no. 5, pp. 679–688, Oct. 2012, doi: 10.1016/j.sbi.2012.09.002.

[21] M. Westerfield, *The zebrafish book. A guide for the laboratory use of zebrafish (Danio rerio)*, 4th ed. Univ. of Oregon Press, Eugene., 2000.

[22] M. Ledesma-Terrón, D. Pérez-Dones, D. Mazó-Durán, and D. G. Míguez, "High-throughput three-dimensional characterization of morphogenetic signals during the formation of the vertebrate retina," *bioRxiv*, Apr. 2024, https://doi.org/10.1101/2024.04.09.588672.

[23] J. Glenn, "ultralytics/COCO2YOLO: Improvements." Zenodo, May 11, 219AD. doi: 10.5281/zenodo.2738322.

APPENDIX

The same techniques used on the tibia have been applied to the femur, resulting in 10 additional models evaluated under the same conditions.

Table A1. Summary of the average value of Dice score (Dice), recall (Rec.), and precision (Prec.) for each model

		Non Enhanced			Enhanced		
	Training images	*Dice*	*Rec.*	*Prec.*	*Dice*	*Rec.*	*Prec.*
Femur	3	0.883	0.846	0.978	0.864	0.828	0.979
	15	0.973	0.977	0.969	0.971	0.972	0.970
	41	0.974	0.975	0.973	0.974	0.975	0.973
	83	0.974	0.976	0.973	0.974	0.975	0.973
	124	0.975	0.976	0.974	0.974	0.975	0.973

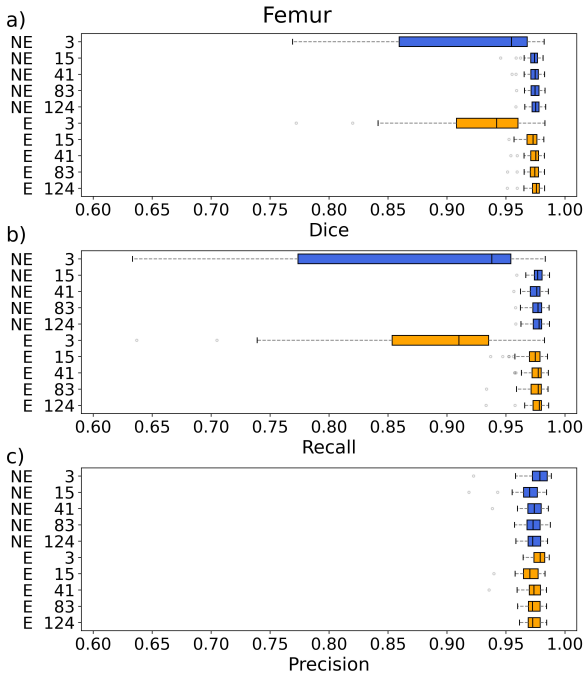
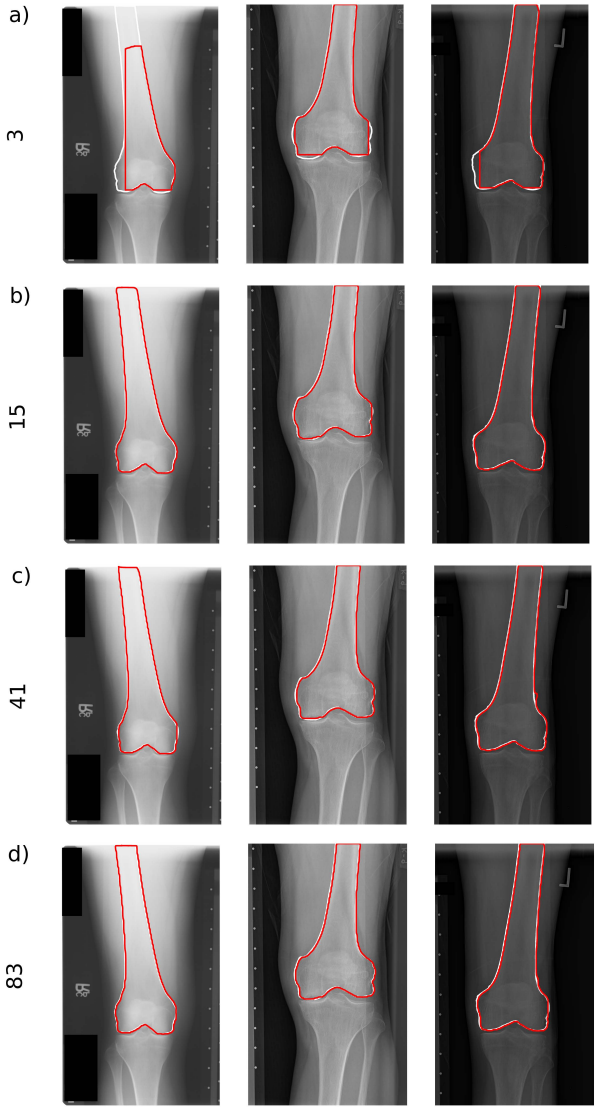


Fig. A1: Graphic summaries of all test images evaluated with all femur models. In blue, models trained with Non-Enhanced (NE) images, while in orange; with them Enhanced (E)

Given the nature of femur segmentations, the problem is more similar to what FastSAM was trained for, the whole segmentation of an object. As shown in Fig. A1, this translates into a very good performance as soon as a few femurs are shown to the model, adjusting it to find this type of object but applying the same strategy as the one used in its original training dataset SA-1B. Also, no significant differences were found between models trained with the same amounts of both types of images (enhanced and non-enhanced). In Table A1, a numerical representation of the performance of each femur model is shown with the average of the dice score distributions shown in Fig. A1. Identically to the case studied for the tibia, the non-enhanced models gradually increase their performance with more train samples, while the enhanced version saturates faster, needing fewer samples to achieve similar results. Probably because of a reduction in the complexity of the problem due to standardization in illumination conditions and visibility overall.

In Fig. A2 examples of the performance of each model are shown for the femur. With 3 patients in the training dataset, the shape is almost completely captured and with 15 training images onwards the predictions get asymptotically better.



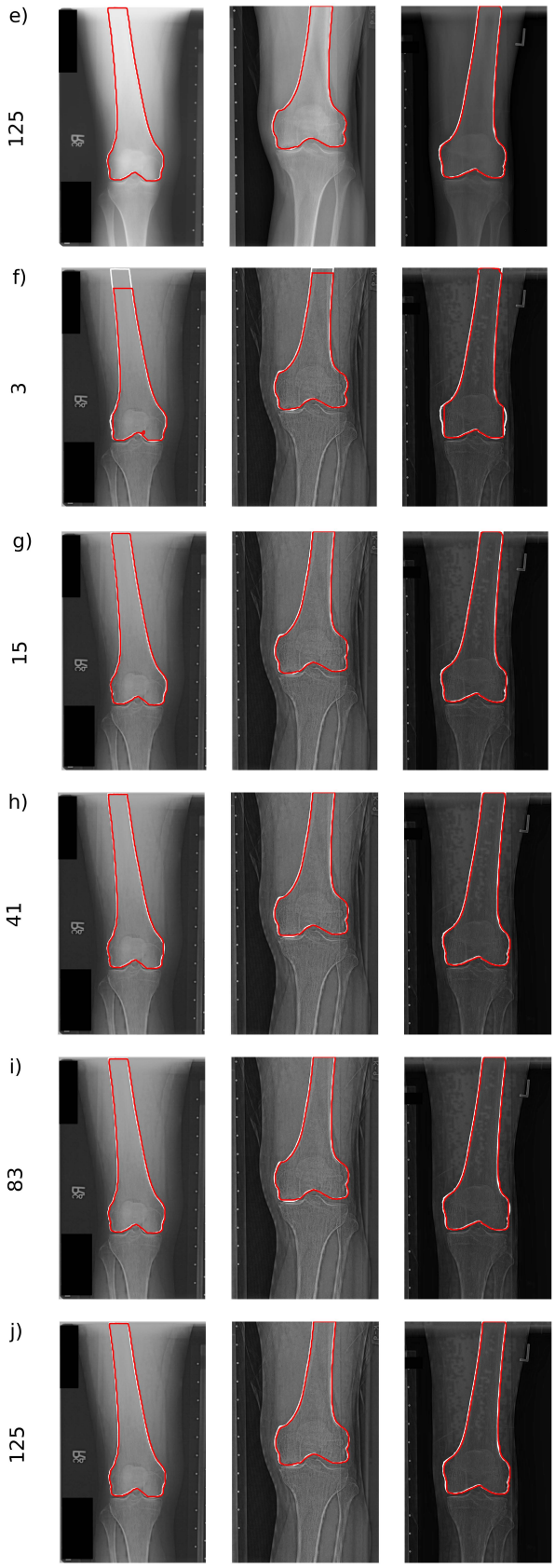


Fig. A2: Contours of ground truth (white) and predictions (red). Rows a), b), c), d), and e) were trained with the original images, while f), g), h), i), and j) with their enhanced counterparts. At the left of each row, the number of images used to finetune the applied model is shown.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Data curation, S.P.-E. and D.P-D.; Formal analysis, S.P.-E., and D.P.-D.; Funding acquisition, F.J.B., J.R.P., G.G.P., D.G.M., and A.P.M.; Investigation, S.P.-E., and D.P.-D.; Methodology, S.P.-E., D.P.-D., and A.P.M.; Project administration, A.P.M.; Resources, S.P.-E., D.P.-D., I.R.-P., N.O.-V., D.G.M. and, A.P.M.; Software, S.P.-E., and D.P.-D.; Supervision, A.P.M.; Validation, S.P.-E., and D.P.-D.; Visualization, S.P.-E.; Writing—original draft, S.P.-E., and D.P.-D.; Writing—review and editing, A.P.M. All authors have read and agreed to the published version of the manuscript

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

We acknowledge financial support under grant PID2022-138322OB-100 funded by MCIN/AEI/ and by "ERDF A way of making Europe". Also, Xunta de Galicia funded this research under Research Grant No. 2021-PG036. We also acknowledge the research network RED2022-134573-T as well, funded by Ministerio de Ciencia e Innovación (MCIN/AEI/10.13039/501100011033) and by 'ERDF: A way of making Europe', by the European Union. And finally; the project PMPTA22/00115 DEL ISCIII, MADRID SPAIN.

Conflict of Interest

We declare no conflict of interest.