SAMUSA: Segment Anything Model 2 for UltraSound Annotation

Baptiste Podvin[1,3], Toby Collins[1,2], Güinther Saibro[1], Chiara Innocenzi[1], Yuchuan Yang[1], Flavio Milana[1,4], Yvonne Keeza[2], Grace Ufitinema[2], Florien Ujemurwego[2], Guido Torzilli[4], Jacques Marescaux[1,2], Daniel George[3⋆], and Alexandre Hostettler[1,2⋆]

[1] Ircad France, Strasbourg, France
[2] Ircad Africa, Kigali, Rwanda
[3] Icube, University of Strasbourg, France
[4] IRCCS Humanitas Research Hospital, Milan, Italy
baptiste.podvin@ircad.fr

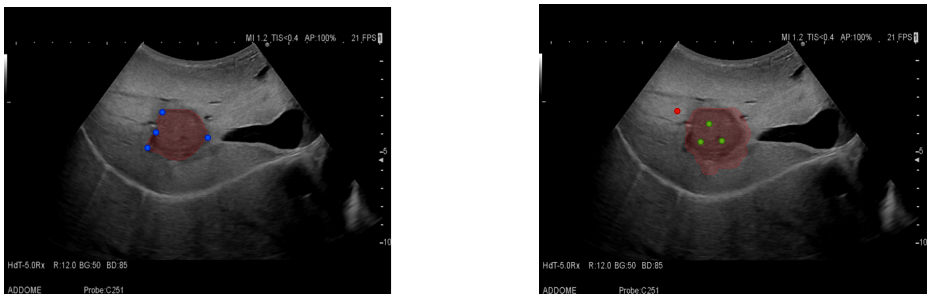
Abstract. Interactive segmentation tools, such as SAM2, have shown strong performance in reducing annotation effort in natural images. However, unlike natural images, ultrasound images and videos often lack well-defined structure boundaries, which significantly degrade the performance of region-based point prompts in SAM models. To address these limitations, we introduce the Segment Anything Model 2 for UltraSound Annotation (SAMUSA). SAMUSA is based on SAM2 and introduces a new prompt strategy with boundary and temporal points, along with a novel boundary loss function, enabling the model to more efficiently segment structures with poorly defined boundaries, such as liver masses. We integrated SAMUSA as a 3D Slicer plugin, where it can be used for US videos and 3D US volumes segmentation. We present a prospective user study involving 6 participants (3 surgeons and 3 radiographers), which showed an average 34.1% annotation time reduction for image liver mass segmentation.

Keywords: SAM2 · Ultrasound · Video annotation · Boundary prompts

1 Introduction and Background

Ultrasound (US) is a major imaging modality in medical diagnosis and image-guided surgery, however, it presents challenges due to noise, operator-dependence, and imaging artifacts. Deep learning models, and especially ultrasound segmentation models have been proposed to assist image interpretation, structure recognition, and measurement. State-of-the-art models are trained on manual segmentations from clinical experts, however, this is often costly and time-consuming, presenting an important clinical translation barrier. Interactive, AI-assisted segmentation models, such as SAM [7] have the potential to reduce this barrier by generating segmentation masks from simple user interactions (prompts), like

⋆ These authors contributed equally to this work.



(a) (b)

Fig. 1: Comparison of lesion segmentation with point prompts: SAMUSA (a) uses boundary points, compared to SAM2 (b), which uses region-based points with two labels (positive/inside, and negative/outside).

bounding boxes or point clicks. SAM was originally proposed for non-medical image segmentation, however, it has been shown to generalize poorly to medical image domains. Recent works such as MedSAM [10], and SAMed [20] overcome this limit with domain-specific fine-tuning. Recently, SAM2 [15] was proposed; an extension of SAM to handle video data using temporal segmentation propagation. Several works propose to fine-tune SAM2 to handle medical imaging modalities, including CT by treating slices along a principal axis as video frames [11, 19, 22]. SAM2 has not been considered explicitly for US data in the literature.

We propose SAMUSA (Segment Anything Model 2 for US Annotation), an AI-assisted annotation tool that extends SAM2 for interactive AI-assisted US segmentation. Our main scientific contribution involves identifying and addressing practical limitations in its prompting mechanism. Specifically, SAM, SAM2, and their previous adaptations for medical image segmentation use point-click prompts, termed region-based points, involving clicking points that are either outside or inside the structure of interest (Fig. 1(b)). While this approach works well for structures with clear boundaries, US images often have ambiguous boundaries. In these cases, segmentation depends on the user's clinical knowledge and contextual cues from other video frames. We hypothesize that a different type of prompt—boundary point prompts (Fig. 1(a)), where users click on where they believe the structure's boundary lies—is more effective than region point prompts because it directly encodes the user's mental decision about the boundary. Additionally, we introduce temporal prompts to mitigate misclassification issues in SAM2 and reduce inference time when processing long video sections where the structure is not visible. We experimentally validate our work in a user study involving clinicians and medical image annotation experts. The study shows that, when fine-tuned on US data, SAM2 can work well in certain scenarios, however, it performs significantly worse in zero-shot applications for hard-to-segment structures (liver masses), compared to SAMUSA.

Table 1: Models were trained on public datasets with static US images (sizes given in terms of number of images), and US videos (sizes given in terms of number of videos). LV - left ventricle, LA - left atrium, Myo. - myocardium.

Data Type	Dataset	Train size	Val size	Test size	Segmented structure(s)
Static 2D Image	Liver [3]	514	73	148	Liver
	FASS [1]	1588	-	-	Fetal abdominal
	BUSBRA [5]	1312	187	376	Breast
	TG3K [6]	3585	-	-	Thyroid gland
	TN3K [6]	2303	575	614	Thyroid nodule
	Nerve[13]	2091	232	-	Brachial Plexus
	OTU [21]	1028	146	295	Ovarian tumors
	STMUS NDA [12]	-	-	3484	Muscle
	Trusted [14]	-	-	3697	Kidney parenchyma
	IUSLL (private)	-	-	35 573	Liver mass
Video	SegThy [8]	32	-	-	Thyroid
	CAMUS [9]	700	100	200	Myo.,LV,LA
	Thyroid cine clip [17]	134	19	39	Thyroid
	Trusted	46	-	13	Kidney parenchyma
	IUSLL (private)	-	-	426	Liver mass

2 Methods

2.1 SAMUSA overview

SAMUSA builds upon the neural network architecture introduced by SAM2 [15], retaining its key components: the Image Encoder, Prompt Encoder, Memory Mechanism, and Mask Decoder. SAMUSA introduces two novel prompt mechanisms: boundary and temporal points. The first mechanism, called **boundary points prompts**, replaces the region-based points prompts used in SAM2. These boundary points are integrated into the loss functions as $\mathcal{L}_{\text{SAMUSA}} = \mathcal{L}_{\text{SAM2}} + \lambda\mathcal{L}_{\text{boundary}}$, where $\mathcal{L}_{\text{SAMUSA}}$ represents the loss function used to train our model, combining the original SAM2 loss, $\mathcal{L}_{\text{SAM2}}$, with our novel boundary loss, $\mathcal{L}_{\text{boundary}}$. The weighting factor $\lambda = 2$ controls the contribution of the boundary loss to the overall objective.

2.2 Boundary prompt loss and training point sampling

Our boundary prompt loss is inspired by Roth *et al.* [16], which penalizes a boundary point being located far from the border of a predicted segmentation mask M. The loss is computed in three steps. First, M is blurred using mean filtering (7×7 kernels), repeated N times (we use a default of $N = 20$), generating a smoothed normalized mask $G_I(x, y)$. Second, a heatmap $G_P(x, y)$ is generated for the boundary points, created by applying a Gaussian kernel centered at each boundary point ($\sigma = 2$). The final heatmap is obtained by summing the Gaussians of all boundary points. The resulting loss is then computed as:

$$\mathcal{L}_{\text{boundary}} = \exp\left(-\sum_{x,y} G_I(x, y) \cdot G_P(x, y)\right) \tag{1}$$

During training, we propose the following mechanism to sample boundary points from batches of training images, with their associated segmentation masks. Boundary points are generated in a manner inspired by Dupont et al. [2]. First, we choose a random point on the mask's boundary associated with the training image. We then select a second point on the boundary furthest from the first point. To simulate the effect of a user adding boundary points to refine the segmentation, additional boundary points are iteratively sampled (those furthest from the predicted segmentation boundary). To improve variability, 2 images per batch are randomly chosen for corrections, with a random number of sampled points varying from 2 to 10.

2.3 Temporal points

In ultrasound videos, a structure can come in and out of view as the operator sweeps the probe over the structure. The visual appearance of a structure during a probe sweep can vary significantly, often cause failures in SAM2's classification head (failure to recognize frames for which the structure is visible). To resolve this issue, we introduce temporal prompts, where a user defines start/end visibility time windows. We use these temporal prompts in SAMUSA for the dual purpose of additional robustness (by bypassing the classification head of the mask decoder), and also for faster inference, where we do not perform mask propagation in frames beyond start/end times.

2.4 Datasets and implementation details

We sourced various public US image and video datasets as described in Table 1. These datasets were combined into a training, validation, and test super-dataset, using the splits as proposed by each public dataset. In addition, a private retrospective video dataset was collected from our partner Humanitas Research hospital, comprising anonymous intra-operative US videos from 59 patients undergoing open liver lesion resection. This dataset was used only for testing purposes and is referred to as IUSLL (Intra-operative US Liver Lesions). We trained SAMUSA and baseline methods on the same datasets, naming the retrained SAM2 model SAM2-US. Before the user video study, we found that SAM2-US model was not a practical baseline due to the excessive time required for mask propagation in frames where the structure was not visible, and frequent classification errors. To address this, we incorporated the temporal point mechanism, referring to the updated baseline as SAM2-US^{TP}. All models were trained using one Nvidia A100, Python 3.10, and PyTorch 2.1. For the user study, we programmed a simple user interface as a 3D Slicer [4] plugin (version 2.6.2).

3 Experimental validation

3.1 User studies

Cohort and metrics. We recruited six users in this study, three were abdominal surgeons (Users 1-3), and three were radiographers and professional medical

Table 2: Comparison of mean per-image segmentation time and Dice scores between SAMUSA and SAM2-US in image user evaluation for the supervised Trusted and the zero-shot IUSLL dataset. Statistically significant best values are in bold

Task	User	SAM2-US		SAMUSA		Intra-user P-value		Inter-user P-value	
		Time (s)	Dice	Time (s)	Dice	Time	Dice	Time	Dice
Supervised (Trusted)	1	3.99	0.94	4.08	0.94	0.56	0.85	0.33	0.092
	2	**4.1**	0.94	5.7	0.96	4.7×10^{-5}	0.26		
	3	4.14	0.94	4.32	0.96	0.54	0.043		
	4	10.51	0.93	**8.36**	0.93	2.1×10^{-4}	0.88		
	5	6.31	0.95	5.68	0.95	0.24	1.0		
	6	16	0.94	**8.42**	0.95	3.2×10^{-12}	0.04		
Zero-shot (IUSLL)	1	6.77	0.81	**4.82**	0.83	1.0×10^{-4}	0.68	4.4×10^{-3}	0.57
	2	14.06	0.80	**7.1**	0.81	6.3×10^{-10}	0.98		
	3	14.8	0.83	**9.9**	0.81	1.7×10^{-7}	0.48		
	4	25.31	0.82	**15.52**	0.83	1.0×10^{-5}	0.82		
	5	13.15	0.84	**9.7**	0.84	2.2×10^{-8}	0.86		
	6	26.29	0.84	**18.86**	0.84	5.4×10^{-5}	0.93		

image annotators (Users 4-6, each with over 7 years of experience in medical image analysis). Two user studies were conducted: *Image segmentation*, which involved segmenting kidneys from the Trusted dataset and liver lesions from IUSLL in individual US images, and *Video segmentation*, which involved the same structures and datasets in US videos. To effectively introduce users to the annotation tools, we provided user training, an annotation protocol, and various sample images and videos distinct from the final test set, used in a warm-up period. We evaluated both methods using time and Dice score metrics. Reference segmentations, from which a user's Dice scores were assessed, were established from all users' segmentations using the STAPLE algorithm [18].

Image segmentation. Each user performed two tasks: (1) kidney segmentation using images and (2) liver lesion segmentation. For each task, 100 random test images were selected, ensuring an equal number of images per patient to reduce bias. Users annotated each image using two models: SAMUSA and SAM2-US. To minimize familiarity bias, half of the test images were segmented using SAMUSA first, while the other half were segmented with SAM2-US first. Due to distribution non-normality, intra-user time and Dice difference were statistically assessed with paired Wilcoxon signed-rank tests. Inter-user average time and Dice differences were assessed with a paired t-test. The average time to segment each image, average DICE scores, and p-values are shown in Table 2.

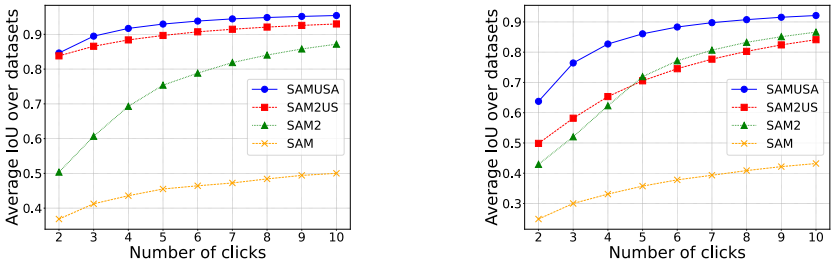
In the zero-shot task, there was a significant statistical difference favoring SAMUSA. Our model outperforms SAM2-US for all users, being 34.1% faster on average. In the supervised task, there was no statistical difference in the time difference for all users. Dice scores show strong agreement among users in the supervised setting, but less agreement in the zero-shot setting, as lesions can

Table 3: Comparison of mean per-video segmentation time and Dice scores between SAMUSA and SAM2-US^{TP} in video user evaluation for the supervised Trusted and the zero-shot IUSLL dataset. Statistically significant best values are in bold

Task	User	SAM2-US^{TP}		SAMUSA		Intra-user P-value		Inter-user P-value	
		Time (s)	Dice	Time (s)	Dice	Time	Dice	Time	Dice
Supervised (Trusted)	User 1	80.1	0.74	80.1	0.74	1	1	0.14	0.33
	User 2	100.8	0.90	119.1	0.91	0.92	0.62		
	User 3	73.1	0.92	56.4	0.93	0.23	0.49		
	User 4	212.9	0.88	188.7	0.93	0.55	0.02		
	User 5	269.9	0.91	**183.9**	0.92	0.037	0.37		
	User 6	260.2	0.95	219.9	0.93	0.76	0.23		
Zero-shot (IUSLL)	User 1	–	–	–	–	–	–	0.031	0.84
	User 2	113.9	0.75	54.7	0.71	0.084	0.0059		
	User 3	117.6	0.76	89.1	0.74	0.92	0.62		
	User 4	260.3	0.80	247.1	0.87	0.69	0.19		
	User 5	157.8	0.83	139.7	0.83	0.37	0.70		
	User 6	191.9	0.84	174.9	0.81	0.92	0.28		

be more difficult and ambiguous to segment. However, there was no statistical difference between the models. Users also filled out questionnaires and rated the SAMUSA model more favorably, with an average score of 4.5/5 compared to 3.0/5 for SAM2-US (higher is better). SAMUSA required less cognitive load, rated on average 1.8/5 compared to 3.5/5 for SAM2-US (lower is better).

Video evaluation To evaluate SAMUSA's performance presented in Table 3, we conducted a controlled video user study where we selected 10 random kidney videos and 10 random liver lesion videos. The mean number of images per video for Trusted was 707.2 and 159.9 for IUSLL. In this setting, they had to first define the start and end points with temporal prompts and keep these points when switching models. Users selected a frame in a video, segmented it, propagated the mask through the remaining frames, and corrected the prediction if necessary. The user evaluation results in Table 3 suggest that while SAMUSA showed slight advantages in mean time for both tasks, statistical significance was not found, likely due to the small sample size. However, we demonstrate significant inter-user time saving, favoring SAMUSA in the liver lesion task. Users also filled out questionnaires and rated the SAMUSA model higher, giving it a score of 4.1/5 compared to 3/5 for SAM2-US^{TP}. They also reported that SAMUSA required less cognitive load (the less the better), rating it on average 2.3/5 compared to 3/5 for SAM2-US^{TP}. Finally, users have noted that SAMUSA is more robust than SAM2-US^{TP}. When using positive and negative points, segmentation can be less precise, as these points continuously add or remove regions without explicitly defining the object's border. In contrast, boundary points directly delineate the structure, leading to more accurate segmentation.



(a) Supervised (b) Zero shot

Fig. 2: Mean IoU comparison over 5 datasets when SAMUSA and SAM2-US are supervised (a) and over 3 datasets in zero-shot (b)

Table 4: Average Dice/IoU over 10 points on images. (*) denotes models tested in zero-shot setting and ($^+$) denotes models tested in supervised setting

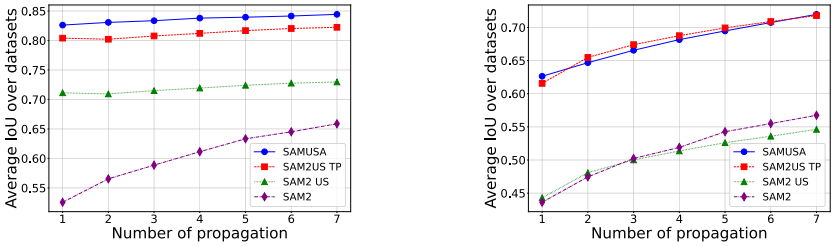
Dataset	SAM*	SAM2*	SAM2-US*	SAMUSA*	SAM2-US$^+$	SAMUSA$^+$
Liver	0.50/0.39	0.71/0.59	–	–	0.93/0.87	**0.95**/**0.91**
BUSBRA	0.76/0.66	0.91/0.85	–	–	0.95/0.91	**0.96**/**0.92**
TN3K	0.71/0.59	0.87/0.79	–	–	0.94/0.89	**0.95**/**0.92**
OTU	0.52/0.40	0.86/0.78	–	–	0.95/0.91	**0.97**/**0.94**
Trusted	0.34/0.22	0.80/0.70	0.81/0.70	**0.91**/**0.87**	0.94/0.89	**0.95**/**0.91**
STMUS	0.45/0.60	0.81/0.71	0.82/0.71	**0.90**/**0.83**	–	–
IUSLL	0.41/0.54	0.82/0.71	0.83/0.71	**0.90**/**0.83**	–	–

3.2 Simulation studies

As part of our ablation study, we compare SAMUSA's single-image and video performance with SAM variants, each utilizing simulated boundary and region point prompts, respectively.

Image segmentation We compare the performance of SAMUSA and SAM variants using single-image simulated points in both supervised and zero-shot settings. For the SAM variants, the first point prompt was placed at the center of the ground truth mask, and subsequent points were iteratively positioned in the areas with the largest errors, while SAMUSA used the same approach as in training. Table 4 presents the results for all methods using 10 click prompts across multiple datasets, while Figure 2 shows the mean IoU across these datasets for varying numbers of clicks. As shown in Figure 2, SAMUSA outperformed all SAM variants for any number of clicks. For example, it surpasses SAM2-US and SAM2 by an average of 10% and 15% on the zero-shot task with two clicks.

Video simulations We performed a coarse simulation of video prompting, based on the observation that users often made corrective prompts as soon as mask propagation deviated unacceptably from their desired segmentation. We simulated point prompts for the first video frame containing a mask, using the



(a) Supervised (b) Zero shot

Fig. 3: Mean IoU comparison over 2 datasets when SAMUSA and SAM2-US are supervised (a) and over 2 video datasets in zero-shot (b).

Table 5: Average Dice/IoU over 8 propagations on videos.(*) denotes models tested in zero-shot setting, ($^+$) denotes models tested in supervised setting ($^{\text{TP}}$) denotes applying temporal prompts.

Model	CAMUS	Thyroid	Trusted	IUSLL
SAM2*	0.90/0.82	0.70/0.61	0.43/0.39	0.73/0.64
SAM2-US$^+$	0.95/0.91	0.82/0.74	0.65/0.50	–
SAM2-US$^{\text{TP}}$ $^+$	0.95/0.91	0.82/0.74	0.82/0.77	–
SAMUSA$^+$	**0.96**/**0.92**	**0.84**/**0.76**	**0.87**/**0.83**	–
SAM2-US*	–	–	0.46/0.41	0.70/0.60
SAM2-US$^{\text{TP}}$*	–	–	0.74/0.69	**0.77**/**0.67**
SAMUSA*	–	–	**0.75**/**0.70**	0.76/**0.67**

single-image algorithm explained earlier, with 5 clicks. The predicted mask was then propagated to subsequent frames until an incorrect segmentation mask was predicted, which we defined as having an IoU score below 0.75. For that frame, the same number of new clicks was simulated, followed by mask propagation. We repeated this process up to 7 times.

In Figures 3a and 3b, we show that SAMUSA outperforms all models in the supervised setting and demonstrates the performance gain from including temporal points. Table 5 shows the average performance across all datasets. However, unlike our user study with human annotators, no significant difference was observed between SAMUSA and SAM2-US$^{\text{TP}}$. It is important to note, however, that the simulation was limited to a very coarse approximation of real user interactions.

4 Conclusion

SAMUSA represents a significant advancement in AI-assisted ultrasound annotation, addressing the challenges posed by ambiguous structure boundaries in ultrasound images and videos. By integrating boundary and temporal prompts, SAMUSA enhances segmentation accuracy and efficiency compared to SAM2, particularly in zero-shot applications where traditional region-based prompts

struggle. The user study demonstrated notable improvements in annotation speed and user experience, reinforcing SAMUSA's practical utility in segmentation annotation. Looking ahead, future research will explore integrating data from other videos or imaging modalities to enhance the segmentation of low-quality ultrasound images and videos for more accurate annotation.

Acknowledgments. We would like to acknowledge the AI engineers at IRCAD Africa — Josiane Uwineza, Jean De Dieu Niyonteze, and Cyriaque Zirimwabagabo — for their assistance with dataset management and user study organization, and the Region Grand Est of France for their support in funding the project.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Da Correggio, K.S., Noya Galluzzo, R., Santos, L.O., Soares Muylaert Barroso, F., Zimmermann Loureiro Chaves, T., Sherlley Casimiro Onofre, A., von Wangenheim, A.: Fetal abdominal structures segmentation dataset using ultrasonic images (2023). https://doi.org/10.17632/4gcpm9dsc3.1
2. Dupont, C., Ouakrim, Y., Pham, Q.C.: Ucp-net: Unstructured contour points for instance segmentation. 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC) pp. 3373–3379 (2021). https://doi.org/10.1109/SMC52423.2021.9658754
3. Egger, J.: 100+ 2d us images and tumor segmentation masks (12 2018). https://doi.org/10.13140/RG.2.2.36586.77761
4. Fedorov, A., Beichel, R.R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D.L., Fennessy, F.M., Sonka, M., Buatti, J.M., Aylward, S.R., Miller, J.V., Pieper, S.D., Kikinis, R.: 3d slicer as an image computing platform for the quantitative imaging network. Magnetic resonance imaging **30 9**, 1323–41 (2012). https://doi.org//10.1016/j.mri.2012.05.001
5. Gómez-Flores, W., Gregorio-Calas, M.J., de Albuquerque Pereira, W.C.: Bus-bra: A breast ultrasound dataset for assessing computer-aided diagnosis systems. Medical physics (2023). https://doi.org//10.1002/mp.16812
6. Gong, H., Chen, J., Chen, G., Li, H., Li, G., Chen, F.: Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. Computers in Biology and Medicine **155**, 106389 (2023). https://doi.org//10.1016/j.compbiomed.2022.106389
7. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.B.: Segment anything. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 3992–4003 (2023). https://doi.org/10.1109/ICCV51070.2023.00371
8. Krönke, M., Eilers, C., Dimova, D., Köhler, M., Buschner, G., Mirzojan, L., Konstantinidou, L., Makowski, M.R., Nagarajah, J., Navab, N., Weber, W., Wendler, T.: Tracked 3d ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. PLoS ONE **17** (2021). https://doi.org/10.1371/journal.pone.0268550

9. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., Lartizien, C., D'hooge, J., Løvstakken, L., Bernard, O.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Transactions on Medical Imaging **38**, 2198–2210 (2019). https://doi.org/10.1109/TMI.2019.2900516
10. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024). https://doi.org//10.1038/s41467-024-44824-z
11. Ma, J., Kim, S., Li, F., Baharoon, M., Asakereh, R., Lyu, H., Wang, B.: Segment anything in medical images and videos: Benchmark and deployment. ArXiv **abs/2408.03322** (2024)
12. Marzola, F., van Alfen, N., Doorduin, J., Meiburger, K.M.: Deep learning segmentation of transverse musculoskeletal ultrasound images for neuromuscular disease assessment. Computers in biology and medicine **135**, 104623 (2021). https://doi.org/10.1016/j.compbiomed.2021.104623
13. Montoya, A., Hasnin, kaggle446, shirzad, Cukierski, W., yffud: Ultrasound nerve segmentation. https://kaggle.com/competitions/ultrasound-nerve-segmentation (2016)
14. Ndzimbong, W., Fourniol, C., Themyr, L., Thome, N., Keeza, Y., Sauer, B., Piéchaud, P.T., Méjean, A., Marescaux, J., George, D., Mutter, D., Hostettler, A., Collins, T.: Trusted: The paired 3d transabdominal ultrasound and ct human data for kidney segmentation and registration research. Scientific Data **12** (04 2025). https://doi.org/10.1038/s41597-025-04467-1
15. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C.K., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R.B., Doll'ar, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. ArXiv **abs/2408.00714** (2024)
16. Roth, H.R., Yang, D., Xu, Z., Wang, X., Xu, D.: Going to extremes: Weakly supervised medical image segmentation. Machine Learning and Knowledge Extraction **3**(2), 507–524 (2021). https://doi.org/10.3390/make3020026
17. S, A.: Thyroid ultrasound cine-clip (2021). https://doi.org/10.71718/7m5n-rh16
18. Warfield, S., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging **23**, 903–921 (2004). https://doi.org/10.1109/TMI.2004.828354
19. Yan, Z., Sun, W., Zhou, R., Yuan, Z., Zhang, K., Li, Y., Liu, T., Li, Q., Li, X., He, L., Sun, L.: Biomedical sam 2: Segment anything in biomedical images and videos. ArXiv **abs/2408.03286** (2024)
20. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. ArXiv **abs/2304.13785** (2023)
21. Zhao, Q., Lyu, S., Bai, W., Cai, L., Liu, B., Wu, M., Sang, X., Yang, M., Chen, L.: A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. CoRR **abs/2207.06799** (2022)
22. Zhu, J., Hamdi, A., Qi, Y., Jin, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2. ArXiv (2024)