CC-SAM: SAM with Cross-feature Attention and Context for Ultrasound Image Segmentation Supplementary Material

Shreyank N Gowda[1] and David A. Clifton[1,2]

[1] Department of Engineering Sciences, University of Oxford, OX3 7DQ Oxford, UK
[2] Oxford Suzhou Centre for Advanced Research, University of Oxford, Suzhou 215123, Jiangsu, China

1 Datasets

- TN3K [5]: This thyroid nodule dataset has 3493 ultrasound images annotated pixel-wise. Collected at Zhujiang Hospital, South Medical University using RESONA 70B, ARIETTA 850, and GE Logiq E9 devices, it's split into 2303 training, 576 validation, and 614 testing images.
- DDTI [12]: Comprising 637 ultrasound images annotated pixel-wise for thyroid nodules, this dataset, collected using TOSHIBA Nemio 30 and TOSHIBA Nemio MX devices, is used as an unseen test set for model generalization evaluations due to its fewer data compared to TN3K.
- TG3K [13]: Initially for segmenting thyroid gland regions in videos, Gong et al. 2023 pre-processed it into a dataset of 3585 images annotated pixel-wise for the thyroid gland. It's now split into 3226 training and 359 validation images.
- BUSI [1]: A dataset of 780 breast ultrasound images with pixel-wise annotations for breast cancer. Sourced from LOGIQ E9 and LOGIQ E9 Agile devices, it classifies images into 487 benign, 210 malignant, and 133 normal cases. With no standard data division, a 7:1:2 ratio is used for training, validation, and testing.
- UDIAT [2]: Containing 163 breast ultrasound images annotated pixel-wise, collected using the Siemens ACUSON Sequoia C512 system. It acts as an unseen test set for model evaluations, sharing the same segmentation focus as BUSI but with less data.
- CAMUS [11]: Designed for 2D echocardiographic assessment, this dataset was sourced from GE Vivid E95 ultrasound scanners at the University Hospital of St Etienne. Covering 500 patients with each offering a 2D apical four-chamber and a two-chamber view sequence, it yields about 18 2D slices per sequence. These contain annotations for the left ventricle (CAMUS-LV), myocardium (CAMUS-MYO), and left atrium (CAMUS-LA). Totaling 19232 images and 57696 masks, a first division yields training-validation and test sets, which are then further split 9:1 for training and validation.
- HMC-QU [3]: A collaboration between Qatar University and Hamad Medical Corporation Hospital resulted in this benchmark echo dataset. It offers 109

4-chamber sequences with pixel-wise myocardium annotations, translating to 2349 2D images. Given its fewer data and similar segmentation focus as CAMUS-MYO, it serves as an unseen test set for model evaluations.

2 Evaluation Metrics

For evaluating the segmentation performance of the CC-SAM method in medical image analysis, the following six metrics are utilized:

1. **Dice Coefficient (Dice):**
$$\text{Dice} = \frac{2 \times TP}{FP + FN + 2 \times TP} \tag{1}$$
where it assesses the similarity between the predicted segmentation and the ground truth.
2. **Hausdorff Distance (HD):** Measures the maximum distance between the predicted segmentation boundaries and the actual boundaries.
3. **Intersection over Union (IoU):**
$$\text{IoU} = \frac{TP}{FP + TP + FN} \tag{2}$$
quantifying the overlap between the predicted and actual segmentations.
4. **Accuracy (ACC):**
$$\text{ACC} = \frac{TN + TP}{TN + TP + FN + FP} \tag{3}$$
reflecting the overall correctness of the segmentation.
5. **Sensitivity (SE):**
$$\text{SE} = \frac{TP}{TP + FN} \tag{4}$$
indicating how well the method detects actual positives.
6. **Specificity (SP):**
$$\text{SP} = \frac{TN}{FP + TN} \tag{5}$$
showing the ability of the method to correctly identify negatives.

In these formulas, TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative, respectively, in the context of segmentation accuracy.

3 Choice of CNN

We experimented with different SOTA models on Imagenet such as ColorNet [8], ResNet [9] and DenseNet [10]. However, in terms of efficiency and performance comparison, ResNet stands out and hence we choose ResNet as our CNN.

4 Augmentation Strategies

We tried several augmentation strategies including mixing strategies [4,7,16] and synthetic sample generation and selection [6, 14, 15]. However, these strategies actually dropped the overall performance of the model. Studying these failure reasons is something that will be considered as future study.

5 Comparison of Different Prompting Strategies

Prompting Strategy	TN3K		BUSI		CAMUS-LV		CAMUS-MYO		CAMUS-LA	
	Dice	HD	Dice	HD	Dice	HD	Dice	HD	Dice	HD
1 pt	83.41	28.11	85.11	27.79	91.89	11.45	84.49	18.31	92.87	11.78
5 pt	84.41	27.89	85.72	26.95	92.71	10.91	84.92	17.24	92.98	11.89
10 pt	83.99	28.15	85.36	27.75	92.74	10.95	84.95	**17.11**	91.71	11.95
Random bounding box	84.45	27.99	85.98	27.11	92.83	10.89	85.19	17.19	93.15	11.41
Grounding DINO	**85.59**	**27.74**	**86.22**	**25.85**	**92.85**	**10.88**	**85.61**	17.11	**93.51**	**11.06**

Table 1: Quantitative comparison of different prompting strategies on seeable US30K data. The performance is evaluated by the Dice score (%) and Hausdorff distance (HD).

The Segment Anything Model has the option for different prompting strategies. Among these include point prompts and bounding box prompts. We compare various prompting strategies in Table 1. We first use varying number of point prompts by randomly sampling points on the image. We then use a random bounding box as input prompt. Finally, we use the proposed manner of Grounding DINO. We see slight improvements by using a random bounding box in comparison to using point prompts. However, the best results are obtained using the proposed strategy.

6 Comparing Prompts for Foundational Models

We also compare with other foundational models by using random point prompts. We use the same set of prompt points and hence this gives us an idea of how the different models perform keeping fixed prompts. We see that the proposed model does much better than all models from a qualitative point of view. These results are shown in Figure 1.

7 Quantitative Results Continued

In the paper, we focus on the Dice score and the HD score when reporting our results. We do a much more detailed analysis by incorporating the other metrics listed. Table 2 lists the results on the TN3K dataset. Similarly results on BUSI dataset can be seen in Table 3, for CAMUS LV dataset can be seen in Table 4 and for CAMUS MYO dataset can be seen in Table 5.

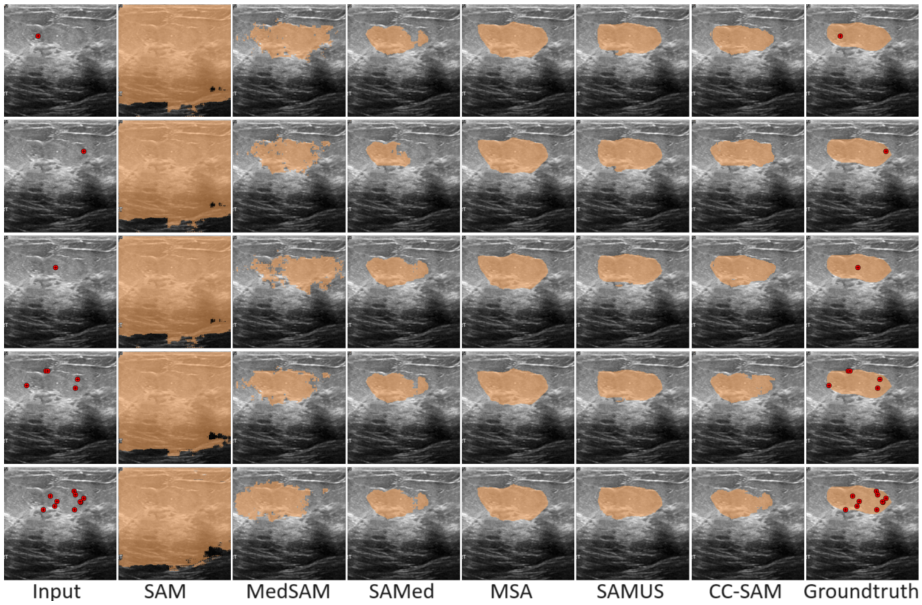

Fig. 1: Comparing how point prompts affect segmentation performance.

Method	Dice%	HD	IoU%	ACC%	SE%	SP%
U-Net	79.01±21.87	34.12±23.77	69.40±23.09	96.44±4.17	84.80±22.70	97.91±2.64
CPFNet	79.43±19.92	33.07±22.71	69.47±22.07	96.56±3.90	83.60±21.43	98.19±2.58
CA-Net	80.52±19.35	33.65±24.93	70.78±21.28	96.41±4.21	86.59±19.90	97.44±4.20
CE-Net	80.37±19.74	32.79±24.28	70.66±21.51	96.40±4.56	85.91±20.22	97.44±4.60
AAU-Net	82.28±18.86	30.53±23.07	73.16±20.80	96.88±4.23	85.68±19.85	98.21±3.38
SwinUnet	70.08±23.29	44.13±25.61	58.19±24.11	94.94±4.35	79.34±24.54	96.55±3.67
SETR	67.80±25.63	44.11±29.10	56.33±26.39	94.35±4.75	76.58±27.34	96.63±4.80
MISSFormer	79.42±21.43	32.85±25.03	69.86±22.92	96.37±4.32	84.39±22.68	97.64±3.86
SAMUS	84.45±15.00	28.22±19.71	75.33±17.60	96.93±4.34	87.13±17.90	97.64±3.08
TransUNet	81.44±19.31	30.98±21.68	72.05±21.08	96.94±3.77	87.13±20.38	97.75±3.08
H2Former	82.48±18.30	30.58±22.10	73.31±20.41	96.95±3.59	85.30±19.34	98.09±2.35
CC-SAM	**85.20±16.93**	**27.10±21.23**	**77.89±19.75**	**97.22±4.58**	**88.56±19.88**	**98.45±3.58**

Table 2: Quantitative comparison results on the seeable TN3K dataset.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Aly, F.: Deep learning approaches for data augmentation and classification of breast masses using ultrasound images. Int. J. Adv. Comput. Sci. Appl **10**(5), 1–11 (2019)
2. Byra, M., Jarosik, P., Szubert, A., Galperin, M., Ojeda-Fournier, H., Olson, L., O'Boyle, M., Comstock, C., Andre, M.: Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network. Biomedical Signal Processing and Control **61**, 102027 (2020)
3. Degerli, A., Kiranyaz, S., Hamid, T., Mazhar, R., Gabbouj, M.: Early myocardial infarction detection over multi-view echocardiography. Biomedical Signal Process-

Method	Dice%	HD	IoU%	ACC%	SE%	SP%
U-Net	78.11±25.45	33.60±32.78	69.60±26.81	96.69±4.71	83.16±23.95	97.95±3.84
CPFNet	80.56±23.77	27.98±28.19	72.28±24.64	97.06±4.74	79.74±25.32	98.83±2.36
CA-Net	81.88±21.55	28.67±28.25	73.49±23.19	96.85±5.31	84.33±22.35	98.27±2.98
CE-Net	81.60±23.53	29.19±31.02	73.62±24.03	96.86±5.47	81.72±23.15	98.09±4.40
AAU-Net	80.81±23.55	28.96±30.06	72.69±25.14	96.76±5.71	82.77±23.96	98.36±4.70
SwinUnet	67.23±25.79	47.02±34.18	55.58±25.88	95.28±5.23	78.94±26.05	96.75±4.97
SETR	68.22±28.28	40.37±32.56	57.68±27.96	95.51±5.98	77.41±29.64	96.61±5.91
MISSFormer	78.43±24.99	33.10±35.08	69.73±25.76	96.26±5.96	83.24±24.08	97.55±5.17
SAMUS	85.77±16.53	25.49±28.10	77.79±18.93	97.47±4.88	87.65±16.48	98.47±4.61
TransUNet	82.22±24.08	27.54±28.25	74.77±24.57	97.26±4.84	83.19±25.18	98.69±3.82
H2Former	81.48±22.91	27.84±27.02	73.34±24.13	96.68±5.50	85.16±22.93	97.71±5.17
CC-SAM	**87.01±18.14**	**24.22±24.59**	**78.49±23.48**	**97.95±4.18**	**89.11±20.47**	**98.89±3.49**

Table 3: Quantitative comparison results on the seeable BUSI dataset.

Method	Dice%	HD	IoU%	ACC%	SE%	SP%
U-Net	93.56±2.43	9.90±4.30	88.00±4.18	98.84±0.70	93.85±3.42	99.36±0.65
CPFNet	93.32±2.37	9.63±3.85	87.57±4.08	98.82±0.62	93.73±3.78	99.36±0.47
CA-Net	93.59±2.13	9.77±3.60	88.02±3.65	98.87±0.58	93.73±3.13	99.41±0.43
CE-Net	93.31±2.46	9.65±3.65	87.55±4.24	98.81±0.66	94.19±3.45	99.31±0.58
AAU-Net	93.32±2.60	9.97±3.92	87.59±4.42	98.82±0.63	94.28±3.33	99.30±0.55
SwinUnet	91.72±2.69	12.80±4.47	84.82±4.45	98.55±0.73	91.34±4.18	99.32±0.58
SETR	92.82±3.10	11.34±4.60	86.76±5.14	98.75±0.65	93.62±4.07	99.28±0.62
MISSFormer	93.25±2.56	9.94±3.84	87.46±4.29	98.82±0.56	94.61±3.29	99.28±0.52
SAMUS	93.73±2.41	9.79±3.67	88.29±4.11	98.89±0.62	94.58±3.07	99.35±0.47
H2Former	93.44±2.30	9.79±3.83	87.78±3.98	98.85±0.60	93.78±3.16	99.38±0.45
CC-SAM	**93.95±2.99**	**9.11±2.58**	**89.56±4.44**	**99.14±0.58**	**94.75±3.89**	**99.41±0.59**

Table 4: Quantitative comparison results on the seeable CAMUS-LV dataset.

Method	Dice%	HD	IoU%	ACC%	SE%	SP%
U-Net	86.86±3.83	16.87±5.99	76.97±5.82	97.27±1.54	87.11±5.88	98.53±0.86
CPFNet	86.68±3.73	16.51±5.58	76.68±5.70	97.27±1.44	86.70±5.37	98.54±0.84
CA-Net	87.21±3.38	16.24±4.80	77.48±5.19	97.38±1.34	87.34±4.83	98.56±0.88
CE-Net	86.47±4.02	16.66±5.76	76.38±6.11	97.23±1.49	86.03±5.72	98.58±0.85
AAU-Net	86.98±3.64	16.49±5.46	77.15±5.53	97.36±1.39	86.29±5.51	98.66±0.82
SwinUnet	84.46±4.00	20.25±6.53	73.30±5.80	96.76±1.64	86.80±5.63	98.00±0.96
SETR	86.20±4.66	18.27±6.77	76.03±6.84	97.15±1.58	86.80±5.43	98.38±1.11
MISSFormer	86.57±3.66	16.50±4.99	76.50±5.53	97.27±1.37	86.02±5.34	98.60±0.84
H2Former	87.31±3.47	16.60±5.17	77.65±5.33	97.37±1.34	88.62±5.16	98.42±0.80
SAMUS	87.46±3.50	16.74±5.48	77.88±5.38	97.37±1.28	89.98±4.26	98.25±1.04
CC-SAM	**88.25±4.11**	**16.11±5.59**	**78.49±23.48**	**97.91±2.11**	**90.12±5.07**	**98.89±1.52**

Table 5: Quantitative comparison results on the seeable CAMUS-MYO dataset.

ing and Control **87**, 105448 (2024). `https://doi.org/https://doi.org/10.1016/j.bspc.2023.105448`
4. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
5. 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)
6. Gowda, S.N.: Synthetic sample selection for generalized zero-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 58–67 (2023)
7. Gowda, S.N., Rohrbach, M., Keller, F., Sevilla-Lara, L.: Learn2augment: learning to composite videos for data augmentation in action recognition. In: European conference on computer vision. pp. 242–259. Springer (2022)
8. Gowda, S.N., Yuan, C.: Colornet: Investigating the importance of color spaces for image classification. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14. pp. 581–596. Springer (2019)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
11. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE transactions on medical imaging **38**(9), 2198–2210 (2019)
12. Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., Romero, E.: An open access thyroid ultrasound image database. In: 10th International symposium on medical information processing and analysis. vol. 9287, pp. 188–193. SPIE (2015)
13. Wunderling, T., Golla, B., Poudel, P., Arens, C., Friebe, M., Hansen, C.: Comparison of thyroid segmentation techniques for 3d ultrasound. In: Medical Imaging 2017: Image Processing. vol. 10133, pp. 346–352. SPIE (2017)
14. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5542–5551 (2018)
15. Ye, J., Xue, Y., Long, L.R., Antani, S., Xue, Z., Cheng, K.C., Huang, X.: Synthetic sample selection via reinforcement learning. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. pp. 53–63. Springer (2020)
16. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019)