고려대학교 정보대학
Korea University
College of Informatics

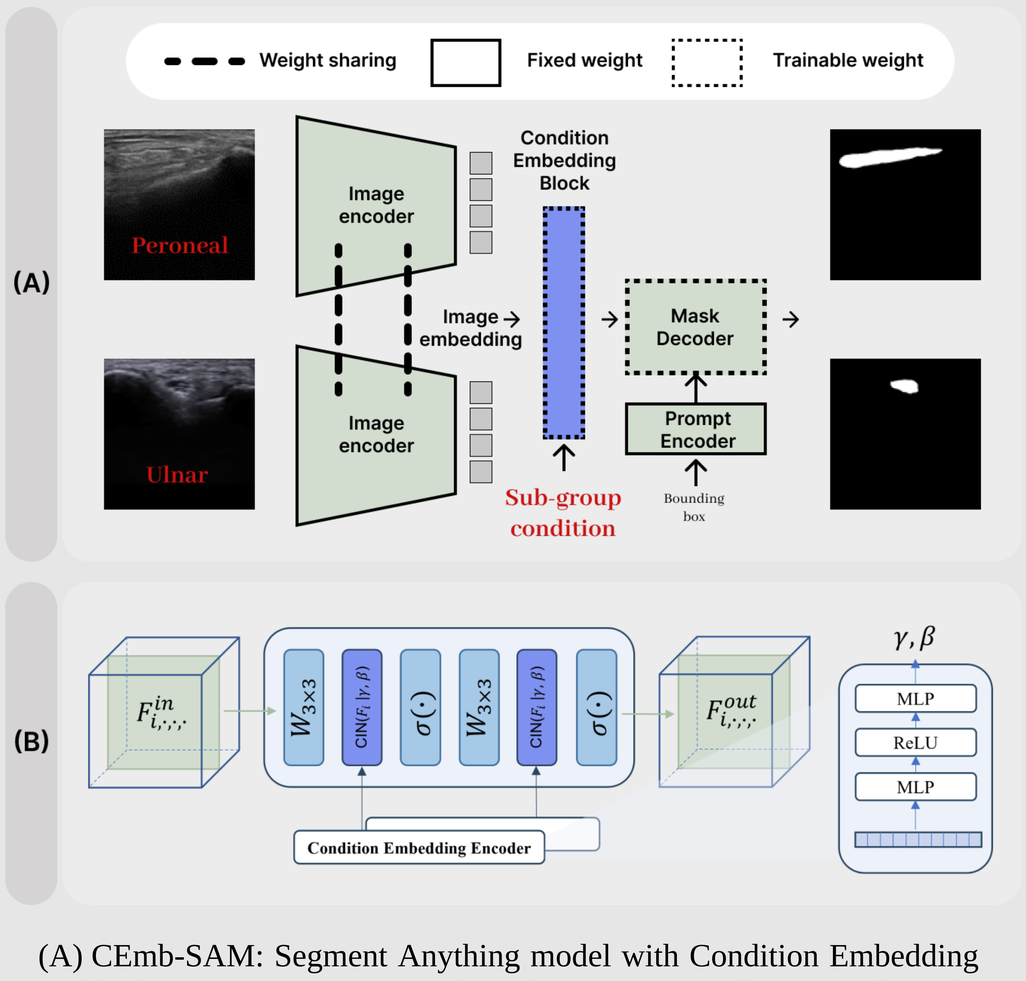
CEmb-SAM: Segment Anything Model with Condition Embedding for Joint Learning from Heterogeneous Datasets

Dongik Shin , Beomsuk Kim, M.D. , and Seungjun Baek

Abstract

Automated segmentation of ultrasound images can assist medical experts with diagnostic and therapeutic procedures. Although using the common modality of ultrasound, one typically needs separate datasets in order to segment, for example, different anatomical structures or lesions with different levels of malignancy. In this paper, we consider the problem of jointly learning from heterogeneous datasets so that the model can improve generalization abilities by leveraging the inherent variability among datasets. We merge the heterogeneous datasets into one dataset and refer to each component dataset as a subgroup. We propose to train a single segmentation model so that the model can adapt to each sub-group. For robust segmentation, we leverage recently proposed Segment Anything model (SAM) in order to incorporate sub-group information into the model. We propose SAM with Condition Embedding block (CEmb-SAM) which encodes sub-group conditions and combines them with image embeddings from SAM. The conditional embedding block effectively adapts SAM to each image sub-group by incorporating dataset properties through learnable parameters for normalization. Experiments show that CEmb-SAM outperforms the baseline methods on ultrasound image segmentation for peripheral nerves and breast cancer. The experiments highlight the effectiveness of CEmb-SAM in learning from heterogeneous datasets in medical image segmentation task

Method



(A) CEmb-SAM: Segment Anything model with Condition Embedding block. Input images come from heterogeneous datasets. The sub-group condition is fed into Condition Embedding block and encoded into sub-group representations. Next, the image embeddings are combined with sub-group representations. The image and prompt encoders are frozen during the fine-tuning of Condition Embedding block and mask decoder.
(B) Detailed description of Condition Embedding Block.

The training dataset is a mixture of m heterogeneous datasets or sub-groups. The training dataset with m mutually exclusive sub-groups $\mathcal{D} = g_1 \cup g_2 \cup \cdots \cup g_m$ consists of N samples $\mathcal{D} = \{(x_i, y_i, y_i^a)_{i=1}^N\}$ where x_i is an input image, y_i is a corresponding ground-truth mask. The sub-group condition $y_i^a \in \{0, \cdots, m-1\}$ represents the index of the sub-group the data belongs to. The peripheral nerve dataset consists of seven sub-groups, six different regions at the peroneal nerve (located below the knee) and a region at the ulnar nerve (located inside the elbow). The BUSI dataset consists of three sub-groups: benign, malignant, and normal. Table 1 shows the detailed description of sub-group indices and variables.

Table 1: Summary of the predefined sub-group conditions of peripheral nerve and BUSI datasets.

Study	Region	Sub-group	m=7	Study	Region	Sub-group	m=3
Nerve	Peroneal	FH	0	BUSI	Breast	Benign	0
		FN	1			Malignant	1
		FN+1	2			Normal	2
		FN+2	3				
		FN+3	4				
		FN+4	5				
	Ulnar	Ulnar	6				

The sub-group condition is encoded into learnable parameters γ and β, and the input feature F^{in} is normalized with given parameters.

$$\gamma = W_2 \cdot \sigma(W_1 \cdot W_\gamma \cdot x_\gamma^a)$$
$$\beta = W_2 \cdot \sigma(W_1 \cdot W_\beta \cdot x_\beta^a)$$

where $W_1, W_2 \in \mathbb{R}^{C\times C}$ are FCNs, $\sigma(\cdot)$ represents ReLU, and x_γ^a, x_β^a are condition representation of a given sub-group. The image embedding is normalized as follows:

$$\mathrm{CIN}(x_i|\gamma,\beta) = \gamma\frac{x_i - \mathrm{E}[x_i]}{\sqrt{\mathrm{Var}[x_i] + \epsilon}} + \beta$$

The proposed Condition Embedding block consists of two independent consecutive CIN layers:

$$\mathrm{F}^{\mathrm{mid}} = \sigma\big(\mathrm{CIN}(W_{3\times3} \cdot x_i|\gamma_1,\beta_1)\big)$$
$$z = \sigma(CIN(\big(W_{3\times3} \cdot \mathrm{F}^{\mathrm{mid}}\big|\gamma_2,\beta_2)))$$

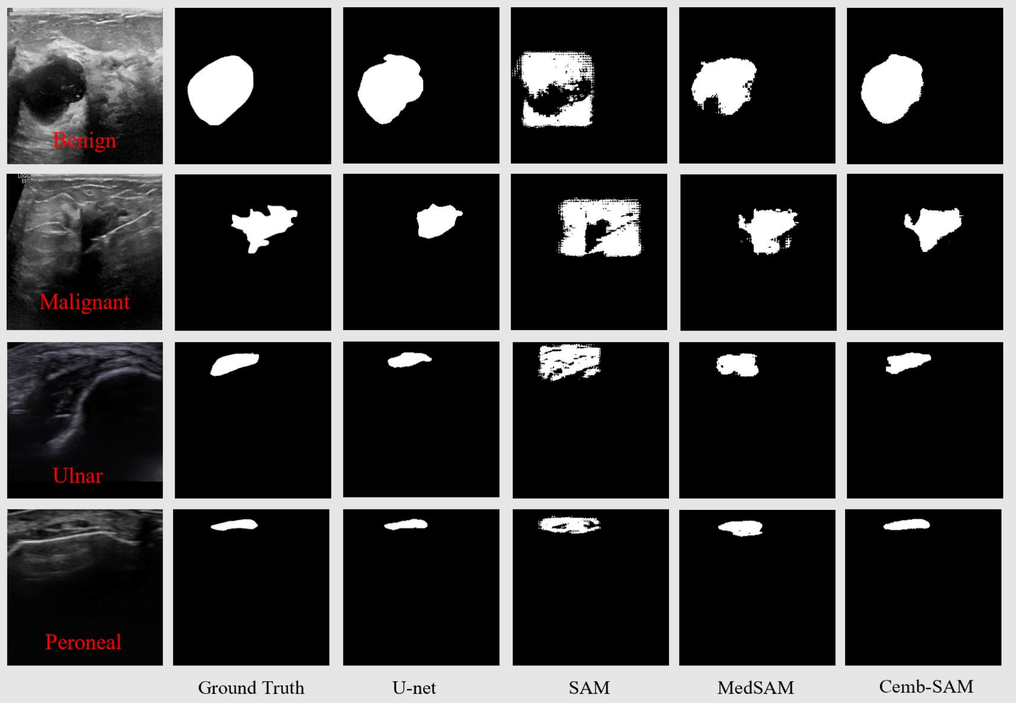
where $\mathrm{F} \in \mathbb{R}^{c\times h\times w}$ represents an intermediate feature map.

Result

Table 2: Performance comparison between U-net, SAM, MedSAM and CEmbSAM on BUSI and Peripheral nerve datasets.

Study	Region	DSC(%)				PA(%)			
		U-net	SAM	MedSAM	Ours	U-net	SAM	MedSAM	Ours
BUSI	Breast	64.87	61.42	85.95	**89.35**	90.72	87.19	90.89	**92.86**
Nerve	Peroneal	69.91	61.72	78.87	**85.02**	92.59	90.58	91.81	**93.90**
	Ulnar	77.04	59.56	83.98	**88.21**	96.49	94.89	96.66	**97.72**

Each dataset was randomly split at a ratio of 80:20 for training and testing. Each training set was also randomly split into 80:20 for training and validation. We used the pre-trained SAM (ViT-B) model as an image encoder. An unweighted sum between Dice loss and cross-entropy loss is used as the loss function. Adam optimizer was chosen to train our proposed method and baseline models using NVIDIA RTX 3090 GPUs.



Segmentation results on BUSI (1st and 2nd rows) and peripheral nerve dataset (3rd and 4th rows)