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

Dongik Shin[1], Beomsuk Kim, M.D.[2], and Seungjun Baek[1(✉)]

[1] Department of Computer Science, Korea University, Seoul, Republic of Korea
{sdimivy014,sjbaek}@korea.ac.kr
[2] Department of Physical and Rehabilitation Medicine, Chung-Ang University College of Medicine, Seoul, Republic of Korea
grit@cau.ac.kr

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 tasks. The code is publicly available at https://github.com/DongDong500/CEmb-SAM

Keywords: Breast Ultrasound · Nerve Ultrasound · Segmentation. · Segment Anything Model

1 Introduction

Image segmentation is an important task in medical ultrasound imaging. For example, peripheral nerves are often detected and screened by ultrasound, which

M. E. Celebi et al. (Eds.): MICCAI 2023 Workshops, LNCS 14393, pp. 275–284, 2023.
https://doi.org/10.1007/978-3-031-47401-9_27

has become a convention modality for computer-aided diagnosis (CAD) [20]. As entrapment neuropathies are considered to be accurately screened and diagnosed by ultrasound [2,3,25], the segmentation of peripheral nerves helps experts identify anatomic structures, measure nerve parameters and provide real-time guidance for therapeutic purposes. In addition, Breast ultrasound images (BUSI) can guide experts to localize and characterize breast tumors, which is also one of the key procedures in CAD [27].

The advancements in deep learning enable an automatic segmentation of ultrasound images, though they still require large, high-quality datasets. The scarcity of the labeled data motivated several studies to propose learning from limited supervision, such as transfer learning [24], supervised domain adaptation [19,22] and unsupervised domain adaptation [6,12,17]. In practice, separate datasets are needed to train a model to segment different anatomical structures or lesions with different levels of malignancy. For example, peripheral nerves can be detected and identified across different human anatomic structures, such as peroneal (located below the knee) and ulnar (located inside the elbow) nerves. Typically, the annotated datasets for peroneal and ulnar nerves are separately constructed, and models are separately trained. However, since the models perform a similar task, i.e., segmenting nerve structures from ultrasound images, one may use a *single* model to be jointly trained with peroneal and ulnar nerves in order to leverage the variability in heterogeneous datasets and improve generalization abilities. A similar argument can be applied to breast ultrasound. A breast tumor is categorized into two types, benign and malignant, and we examine the effectiveness of a single model handling the segmentation of both types of lesions. While a simple approach would be incorporating multiple datasets for training, the characteristics of imaging vary among datasets, and it is challenging to train models which deal with distribution shift and generalize well for the entire heterogeneous datasets [4,26,28].

In this paper, we consider methods to train a single model with heterogeneous datasets jointly. We combine the heterogeneous datasets into one dataset and call each component dataset as a *subgroup*. We consider a model which can adapt to domain shifts among sub-groups and improve segmentation performances. We leverage recently proposed *Segment Anything model* (SAM) which has shown great success in natural image segmentation [14]. However, several studies have shown that SAM could fail on medical image segmentation tasks [5,9,10,16,29]. We adapt SAM to distribution shifts across sub-groups using a novel method for condition embedding, which is called SAM with Condition Embedding block (CEmb-SAM). In CEmb-SAM, we encode sub-group conditions and combine them with image embeddings. Through experiments, we show that the sub-group conditioning guides SAM to adapt to each sub-group effectively. Experiments demonstrate that, compared with SAM [14] and MedSAM [16], CEmb-SAM shows consistent improvements in the segmentation tasks for both peripheral nerves and breast lesions. Our main contributions are as follows:

– We propose CEmb-SAM, which jointly trains a model over heterogeneous datasets leveraging *Segment Anything model* for robust segmentation performances.
– We propose a conditional embedding module to combine sub-group representations with image embeddings, which effectively adapts the Segment Anything Model to sub-group conditions.
– Experiments on the peripheral nerve and the breast cancer datasets demonstrate that CEmb-SAM significantly outperforms the baseline models.

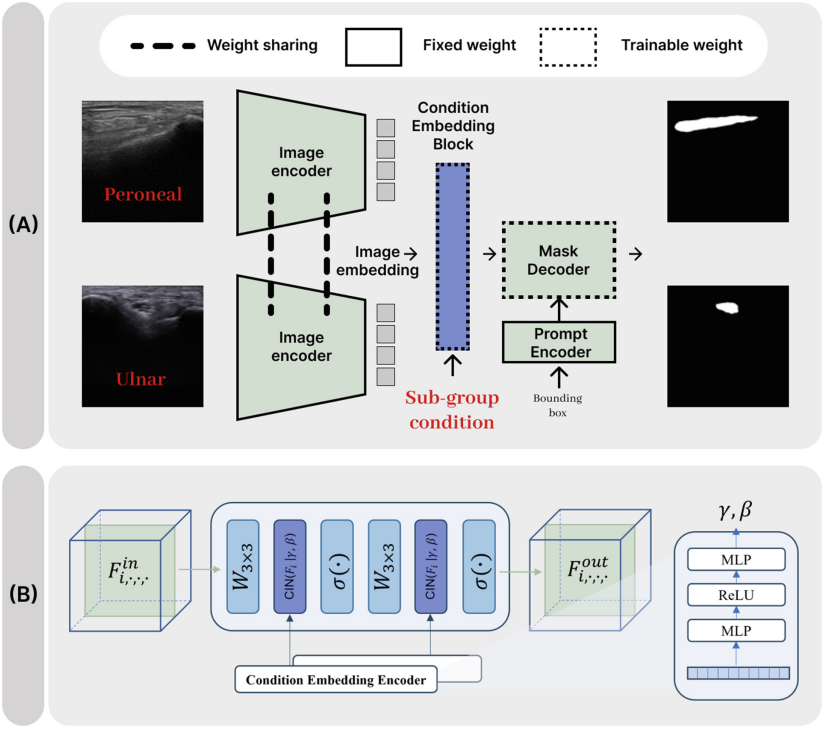

Fig. 1. (A) CEmb-SAM: *Segment Anything model* with Condition Embedding block. Input images come from heterogeneous datasets, i.e., the datasets of peroneal and ulnar nerves, and the model is jointly trained to segment both types of nerves. The sub-group condition is fed into Condition Embedding block 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 sub-group condition is encoded into learnable parameters γ and β, and the input feature F^{in} is scaled and shifted using those parameters.

2 Method

The training dataset is a mixture of m heterogeneous datasets or sub-groups. The training dataset with m mutually exclusive sub-groups $\mathcal{D} = \mathbf{g}_1 \cup \mathbf{g}_2 \cup \cdots \cup \mathbf{g}_m$ 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 $y_i^a \in \{0, \ldots, m-1\}$ 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. The detailed description and sub-group indices and variables are shown in Table 1.

2.1 Fine-Tuning SAM with Sub-group Condition

SAM architecture consists of three components: image encoder, prompt encoder, and mask decoder. Image encoder uses a vision transformer-based architecture [7] to extract image embeddings. Prompt encoder utilizes user interactions, and mask decoder generates segmentation results based on the image embeddings, prompt embeddings, and its output token [14]. We propose to combine sub-group representations with image embeddings from the image encoder using the proposed Condition Embedding block (CEmb). The proposed method, SAM with condition embedding block (CEmb-SAM), uses a pre-trained SAM (ViT-B) model as the image encoder and the prompt encoder. For the peripheral nerve dataset, we fine-tune the mask decoder and CEmb with seven sub-groups. Likewise, we fine-tune the mask decoder on the breast cancer dataset with three sub-groups. The overall framework of the proposed model is illustrated in Fig. 1.

2.2 Condition Embedding Block

We modified the conditional instance normalization (CIN) [8] to combine sub-group representations and image embeddings. Learnable parameters $W_\gamma, W_\beta \in \mathbb{R}^{C \times m}$ where m is the number of sub-groups of the datasets, and C is the number of the output feature maps. A sub-group condition y^a is converted to one-hot vectors, x_γ^a and x_β^a which are fed into Condition Embedding encoder and transformed into sub-group representation parameters γ and β using two fully connected layers (FCNs). Specifically,

$$\gamma = W_2 \cdot \sigma(W_1 \cdot W_\gamma \cdot x_\gamma^a),\ \beta = W_2 \cdot \sigma(W_1 \cdot W_\beta \cdot x_\beta^a) \tag{1}$$

where W_1, $W_2 \in \mathbb{R}^{C \times C}$ are FCN weights, and $\sigma(\cdot)$ represents ReLU activation function.

The image embedding x is transformed into the final representation z using the condition embedding as follows. The image embedding is normalized with mini-batch $\mathcal{B} = \{x_i, y_i^a\}_{i=1}^{N_n}$ of N_n examples as follows:

Table 1. Summary of the predefined sub-group conditions of peripheral nerve and BUSI datasets. FH: fibular head, FN: fibular neuropathy. FN + α represents the measured site is α cm away from the fibular head. m represents the total number of sub-groups.

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

$$\mathrm{CIN}(x_i|\gamma,\beta) = \gamma\frac{x_i - \mathrm{E}[x_i]}{\sqrt{\mathrm{Var}[x_i] + \epsilon}} + \beta \tag{2}$$

where $\mathrm{E}[x_i]$ and $\mathrm{Var}[x_i]$ are the instance mean and variance, and γ and β are given by Condition Embedding encoder. The proposed CEmb consists of two independent consecutive CIN layers with convolutional layers given by:

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

$$z = \sigma(\mathrm{CIN}(W_{3\times3} \cdot F^{\mathrm{mid}}|\gamma_2,\beta_2)) \tag{4}$$

where $F \in \mathbb{R}^{c\times h\times w}$ represents an intermediate feature map, $\mathrm{W}_{3\times3}$ denotes convolution kernel size with 3×3. Figure 1 (B) illustrates the Condition Embedding block.

Table 2. Sample distribution of peripheral nerve and BUSI datasets. FH: fibular head, FN: fibular neuropathy. FN + α represents that the measured site is α cm away from the fibular head.

Dataset	Region	Sub-group	#of samples	Dataset	Region	Sub-group	#of samples
Nerve	Peroneal	FH	91				
		FN	106			Benign	437
		FN+1	77				
		FN+2	58	BUSI	Breast	Malignant	210
		FN+3	49			Normal	133
		FN+4	29				
	Ulnar	Ulnar	1234				
Total			1644	Total			780

Table 3. Performance comparison between U-net, SAM, MedSAM and CEmb-SAM 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**

3 Experiments

3.1 Dataset Description

We evaluate our method on two datasets: (i) a public benchmark dataset, Breast Ultrasound images (BUSI) [1]; (ii) the peripheral nerve ultrasound images collected in our institution. Ultrasound images in the public BUSI dataset are measured from an identical site. The dataset is categorized into three sub-groups: benign, malignant, and normal. The shape of a breast lesion varies according to its type. The benign lesion possesses a relatively round and convex shape. On the other hand, the malignant lesion possesses a rough and uneven spherical shape. The BUSI dataset consists of 780 images. The average image size of the dataset is 500×500 pixels.

The peripheral nerve dataset was created at the Department of Physical Medicine and Rehabilitation, Korea University Guro Hospital. The dataset consists of ultrasound images of two different anatomical structures, the peroneal nerve and the ulnar nerve. The peroneal nerve, on the outer side of the calf of the leg, contains 410 images with an average size of 494×441 pixels. The peroneal nerve images are collected from six different anatomical structures where the nerve stem comes from the adjacent fibular head. FH represents the fibular head, and FN represents fibular neuropathy. FN+α represents that the measured site is α cm away from the fibular head. The ulnar nerve is located along the inner side of the arm and passing close to the surface of the skin near the elbow. The ulnar nerve dataset contains 1234 images with an average size of 477×435 pixels. Table 2 describes the sample distribution of datasets. This study was approved by the Institutional Review Board at Korea University (IRB number: 2020AN0410).

3.2 Experimental Setup

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. SAM comes with three segmentation modes: segmenting everything in a fully automatic way, bounding box mode, and point mode. However, in the case of applying SAM for medical image segmentation, it seems that the segment everything mode is prone to erroneous region partitions. The point-based mode empirically

requires multiple iterations of prediction correction. The bounding box-based mode can clearly specify the ROI and obtain good segmentation results without multiple trials and errors [16]. Therefore, we choose the bounding box prompts as input to the prompt encoder for SAM, MedSAM, and CEmb-SAM. In the training phase, the bounding box coordinates were generated from the ground-truth targets with a random perturbation of 0–10 pixels.

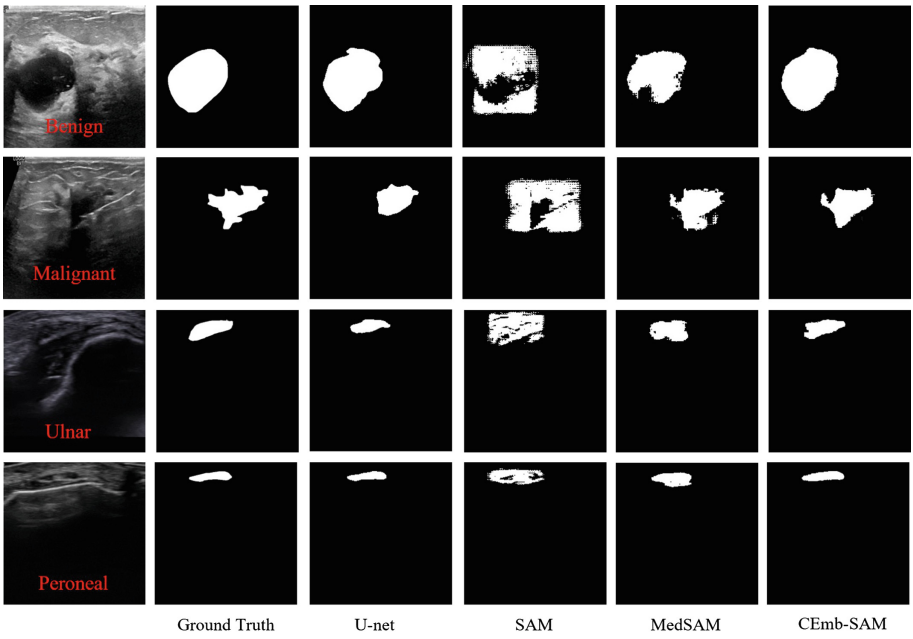

Fig. 2. Segmentation results on BUSI (1st and 2nd rows) and peripheral nerve dataset (3rd and 4th rows).

The input image's intensity values were normalized using Min-Max normalization [21] and resized to $3 \times 256 \times 256$. 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 [11,15]. Adam optimizer [13] was chosen to train our proposed method and baseline models using NVIDIA RTX 3090 GPUs. The initial learning rate of our model is 3e-4.

3.3 Results

To evaluate the effectiveness of our method, we compare CEmb-SAM with the U-net [23], SAM [14], and MedSAM [16]. The U-net is trained from scratch on BUSI and peripheral nerve datasets, respectively. The SAM is used with the bounding box mode. The pre-trained SAM (ViT-B) weights are used as image

encoder and prompt encoder. During inference, the bounding box coordinates are used as the input to the prompt encoder. Likewise, the pre-trained SAM (ViT-B) weights are used as image encoder and prompt encoder in the MedSAM. The mask decoder of MedSAM is fine-tuned on BUSI and peripheral nerve datasets. CEmb-SAM also uses the pre-trained SAM (ViT-B) model as an image encoder and prompt encoder, and fine-tunes the mask decoder on BUSI and peripheral nerve datasets. During inference, the bounding box coordinates are used as the input to the prompt encoder.

For the performance metrics, we used the Dice Similarity Coefficient (DSC) and Pixel Accuracy (PA) [18]. Table 3 shows the quantitative results comparing with CEmb-SAM, MedSAM, SAM (ViT-B), and U-net on both BUSI and peripheral nerve datasets. From Table 3, we observe that our method achieves the best results on both DSC and PA scores. CEmb-SAM outperformed the baseline methods in terms of the average DSC by 18.61% in breast, 14.85% in peroneal, and 14.68% in ulnar, and in terms of the average PA by 3.26% in breast, 2.24% in peroneal and 1.71% in ulnar.

Figure 2 shows the visualization of segmentation results on peripheral nerve dataset and BUSI. The qualitative results show that CEmb-SAM achieves the best segmentation results with fewer missed and false detections in the segmentation of both the breast lesions and peripheral nerves. The results demonstrate that CEmb-SAM is more effective and robust in the segmentation through learning from domain shifts caused by heterogeneous datasets.

4 Conclusion

In this study, we propose CEmb-SAM which adapts the Segment Anything Model to each dataset sub-group for joint learning from the entire heterogeneous datasets of ultrasound medical images. The proposed module for conditional instance normalization was able to guide the model to effectively combine image embeddings with subgroup conditions for both the BUSI and peripheral nerve datasets. The proposed module helped the model deal with distribution shifts among sub-groups. Experiments showed that CEmb-SAM achieved the highest score in DSC and PA on both the public BUSI dataset and peripheral nerve datasets. As future work, we plan to extend our work for improved domain adaptation in which the model is robust and effective under higher degrees of anatomical heterogeneity among datasets.

Acknowledgements. This work was supported by the Korea Medical Device Development Fund grant funded by the Korea Government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 1711195279, RS-2021-KD000009); the National Research Foundation of Korea (NRF) Grant through the Ministry of Science and ICT (MSIT), Korea Government, under Grant 2022R1A5A1027646; the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (No.2021R1A2C1007215); the MSIT(Ministry of Science and ICT), Korea, under the ICT Creative Consilience program(IITP-2023-2020-0-01819) supervised by

the IITP(Institute for Information & communications Technology Planning & Evaluation)

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **28**, 104863 (2020)
2. Beekman, R., Visser, L.H.: Sonography in the diagnosis of carpal tunnel syndrome: a critical review of the literature. Muscle Nerve Official J. Am. Assoc. Electrodiagnostic Med. **27**(1), 26–33 (2003)
3. Cartwright, M.S., Walker, F.O.: Neuromuscular ultrasound in common entrapment neuropathies. Muscle Nerve **48**(5), 696–704 (2013)
4. Davatzikos, C.: Machine learning in neuroimaging: progress and challenges. Neuroimage **197**, 652 (2019)
5. Deng, R., et al.: Segment anything model (SAM) for digital pathology: Assess zero-shot segmentation on whole slide imaging. arXiv preprint arXiv:2304.04155 (2023)
6. Dong, N., Kampffmeyer, M., Liang, X., Wang, Z., Dai, W., Xing, E.: Unsupervised domain adaptation for automatic estimation of cardiothoracic ratio. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 544–552. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_61
7. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. arXiv preprint arXiv:1610.07629 (2016)
9. He, S., Bao, R., Li, J., Grant, P.E., Ou, Y.: Accuracy of segment-anything model (SAM) in medical image segmentation tasks. arXiv preprint arXiv:2304.09324 (2023)
10. Hu, C., Li, X.: When SAM meets medical images: An investigation of segment anything model (SAM) on multi-phase liver tumor segmentation. arXiv preprint arXiv:2304.08506 (2023)
11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
12. Kamnitsas, K., et al.: Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 597–609. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_47
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Kirillov, A., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
15. Ma, J., et al.: Loss odyssey in medical image segmentation. Med. Image Anal. **71**, 102035 (2021)
16. Ma, J., Wang, B.: Segment anything in medical images. arXiv preprint arXiv:2304.12306 (2023)
17. Mahmood, F., Chen, R., Durr, N.J.: Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. IEEE Trans. Med. Imaging **37**(12), 2572–2581 (2018)

18. Maier-Hein, L., Menze, B., et al.: Metrics reloaded: Pitfalls and recommendations for image analysis validation. arXiv:2206.01653 (2022)
19. Motiian, S., Piccirilli, M., Adjeroh, D.A., Doretto, G.: Unified deep supervised domain adaptation and generalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5715–5725 (2017)
20. Noble, J.A., Boukerroui, D.: Ultrasound image segmentation: a survey. IEEE Trans. Med. Imaging **25**(8), 987–1010 (2006)
21. Patro, S., Sahu, K.K.: Normalization: a preprocessing stage. arXiv preprint arXiv:1503.06462 (2015)
22. Redko, I., Morvant, E., Habrard, A., Sebban, M., Bennani, Y.: A survey on domain adaptation theory: learning bounds and theoretical guarantees. arXiv preprint arXiv:2004.11829 (2020)
23. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
24. Shin, H.C., et al.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35**(5), 1285–1298 (2016)
25. Walker, F.O., et al.: Indications for neuromuscular ultrasound: expert opinion and review of the literature. Clin. Neurophysiol. **129**(12), 2658–2679 (2018)
26. Wang, R., Chaudhari, P., Davatzikos, C.: Embracing the disharmony in medical imaging: a simple and effective framework for domain adaptation. Med. Image Anal. **76**, 102309 (2022)
27. Xian, M., Zhang, Y., Cheng, H.D., Xu, F., Zhang, B., Ding, J.: Automatic breast ultrasound image segmentation: a survey. Pattern Recogn. **79**, 340–355 (2018)
28. Zhou, S.K., et al.: A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc. IEEE **109**(5), 820–838 (2021)
29. Zhou, T., Zhang, Y., Zhou, Y., Wu, Y., Gong, C.: Can SAM segment polyps? arXiv preprint arXiv:2304.07583 (2023)