Supervised Contrastive Embedding for Medical Image Segmentation

SANGWOO LEE1,*, YEJIN LEE1,*, GEONGYU LEE1, AND SANGHEUM HWANG1,2
1Department of Data Science, Seoul National University of Science and Technology, Seoul 01811, South Korea
2Department of Industrial and Information Systems Engineering, Seoul National University of Science and Technology, Seoul 01811 South Korea
*Equal contribution

Corresponding author: Sangheum Hwang (e-mail: shwang@seoultech.ac.kr).

This work was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education under Grant 2018R1D1A1A02086017 and Grant NRF-2021R1C1C1011907.

ABSTRACT Deep segmentation networks generally consist of an encoder to extract features from an input image and a decoder to restore them to the original input size to produce segmentation results. In an ideal setting, the trained encoder should possess the semantic embedding capability, which maps a pair of features close to each other when they belong to the same class, and maps them distantly if they correspond to different classes. Recent deep segmentation networks do not directly deal with the embedding behavior of the encoder. Accordingly, we cannot expect that the features embedded by the encoder will have the semantic embedding property. If the model can be trained to have the embedding ability, it will further enhance the performance as restoring from those features is much easier for the decoder. To this end, we propose supervised contrastive embedding, which employs feature-wise contrastive loss for the feature map to enhance the segmentation performance on medical images. We also introduce a boundary-aware sampling strategy, which focuses on the features corresponding to image patches located at the boundary area of the ground-truth annotations. Through extensive experiments on lung segmentation in chest radiographs, liver segmentation in computed tomography, and brain tumor and spinal cord gray matter segmentation in magnetic resonance images, it is demonstrated that the proposed method helps to improve the segmentation performance of popular U-Net, U-Net++, and DeepLabV3+ architectures. Furthermore, it is confirmed that the robustness on domain shifts can be enhanced for segmentation models by the proposed contrastive embedding.

INDEX TERMS Medical image segmentation, contrastive learning, boundary-aware sampling, domain robustness

I. INTRODUCTION

RECENTLY, learning methods based on deep neural networks are rapidly changing the field of medical imaging analysis and are being applied for clinical purposes, e.g., early detection or classification of lesions to improve current practices. Deep neural networks learned from a large amount of data support precise diagnoses and accelerate time-consuming processes requiring medical expertise [1]–[3]. Semantic segmentation in medical imaging analysis is a particularly important area and is essential for diagnosis, monitoring, and treatment [4]. Despite such significant progress, better and reliable performance is required for segmentation models to be widely used in clinical settings. To achieve more precise and reliable segmentation performance, recent studies have proposed several strategies: the design of advanced model architectures [5]–[8] including devising a new information delivery method [9]–[11], or designing a new loss function that is specific to a segmentation task [12]–[18].

Typically, segmentation networks consist of an encoder to extract features (i.e., representations) from an input image and a decoder to generate prediction maps from those extracted features. If a model is ideally trained to meet the original purpose of the encoder and decoder, the feature vectors extracted from the encoder should have the semantic embedding property, i.e., they should be easily distinguishable according to their classes in the corresponding receptive fields. That is, the feature vectors representing the same class of regions should be located close to each other and the feature vectors representing different classes should be
distantly separated in the embedding space as illustrated in Fig. 1. However, we cannot expect that the encoder of extant segmentation models would have this property since previous studies have focused on improving segmentation performance without considering the aforementioned aspect of encoders.

In this study, we propose a novel framework termed supervised contrastive embedding (SCE), which utilizes contrastive loss [19] to encourage the encoder to learn the semantic embedding property during training. Contrastive loss has been actively used for self-supervised learning which aims to learn generic representations from large-scale unlabeled data [20]. To apply contrastive loss to segmentation tasks, we extend the existing contrastive loss by contrasting the feature vectors extracted from the encoder. Each feature vector corresponds to a specific patch in the input image and can be labeled by its position on the ground-truth segmentation mask. Thus, the proposed contrastive loss encourages the encoder to embed closely the feature vectors associated with positive regions and embed them far from the feature vectors associated with negative regions in the embedding space. By contrasting the embedded feature vectors to be easily distinguished in the feature space, the decoder can generate more accurate segmentation results.

Meanwhile, deep neural networks are known to degrade performance when they encounter data from a domain even slightly different from that of the training data [21]. Especially in the field of medical imaging, such situations, termed domain shift, are commonly caused by diverse image acquisition equipment and the personal characteristics of patients. Models are unreliable that fail to generalize to target domains that share semantic information with source domains but have slightly different visual characteristics. Therefore, domain robustness is considered an important research direction along with improvements to model performance [22], [23]. Our SCE contrasts positive and negative regions within different input images to learn the desirable embedding property. Therefore, by utilizing multi-source domains during training, it can be expected to learn domain invariant features.

We comprehensively validated the effectiveness of our proposed method on four different segmentation tasks from diverse imaging modalities, including liver (CT), brain tumor (MR), lung (X-rays), and spinal cord gray matter (MR) segmentation. For the experiments, we adopted model architectures that are widely used in medical image segmentation tasks: U-Net, U-Net++, and DeepLabV3+. Our contributions can be summarized as follows\(^1\):

- We provide qualitative and quantitative results, demonstrating that the encoder of the existing segmentation models does not have the desirable semantic embedding property. In specific, we show that segmentation models learned by the conventional training strategy do not consider the semantic relationship between feature vectors in the embedding space.

- We propose a novel framework called supervised contrastive embedding that utilizes contrastive loss for the encoder to learn the ideal embedding property. Also, we propose an effective sampling strategy called boundary-aware sampling, which further enhances segmentation performance by sampling features from the boundary region of foreground and background. Extensive experiments to demonstrate the advantages of the proposed method were conducted on various medical image datasets and state-of-the-art architectures.

- We also demonstrate that our proposed method is effective in improving the domain robustness of segmentation models when trained on multi-source domains. We validated the effectiveness of our proposed method on domain robustness under a standard domain generalization evaluation protocol.

The remainder of this paper is organized as follows. In Section II, we introduce the related works on semantic segmentation and contrastive learning. In Section III, we provide details on the proposed supervised contrastive embedding with boundary-aware sampling. In Section IV, we present our experimental settings and results for semantic segmentation and domain robustness. Finally, conclusions and discussions are given in Section V.

II. RELATED WORKS

A. SEMANTIC SEGMENTATION

Several methods have been proposed to improve segmentation performance from the perspective of deep neural network architecture. One of the pioneering works is on fully convolutional networks (FCNs) [24]. It successfully outperformed previous segmentation models, and significantly influenced subsequent studies on neural networks-based semantic segmentation. To further enhance FCNs, encoder-decoder architectures that have symmetrical structures were proposed to restore feature maps to the original size of its inputs by a trainable decoder [25], [26]. DeepLab family [5]–[7], [27] is also one of the notable architectural advances utilizing dilated convolutions as a way to increase the size of the receptive field while maintaining the number of trainable parameters. DeepLabV3+ [7], the latest DeepLab series, combines depthwise separable convolutions with dilated convolutions and, thereby, conspicuously reduces the amount of computation while maintaining performance. U-Net [28]

\(^1\)The code is publicly available at https://github.com/daintlab/supervised-contrastive-embedding
and its variants have been proposed for medical image segmentation and gained popularity. Because of its superior performance, U-Net has been widely used for biomedical image segmentation and many subsequent works have adopted the U-shaped architecture as a basic design [29], e.g., UNet++ [9], MultiResUNet [10], TransUNet [11], etc.

Besides advanced architectures, employing a better loss function also effectively improves segmentation performance in deep neural networks. Focal loss [30] assigns weights to the cross entropy loss according to the difficulty of classifying pixels for the model to be more affected by pixels that are hard to classify. Loss functions that aim to optimize the target overlap metrics of segmentation directly, e.g., Dice loss [31], have also been introduced. Salehi et al. [32] proposed a loss function based on the Tversky index and Abraham and Khan [14] introduced Focal Tversky loss, a generalized focal loss with the Tversky index. ClDice [17], [18] is a similarity measure that considers the connectedness of segmented regions, and its differentiable form is used as a loss function. Such pixel-level and region-level losses can be also naturally combined [13], [15]. However, none of the previous approaches explicitly encourages the encoder in segmentation networks to learn the semantic embedding property.

**B. CONTRASTIVE LEARNING**

The operating principle of contrastive learning is to learn representations by contrasting positive pairs against negative pairs. Based on this concept, contrastive loss has been extended in various forms and has achieved remarkable progress in recent years [20], [33]–[35]. For example, SimCLR [20] used contrastive loss to learn generic representations with a large-scale unlabeled dataset in a self-supervised manner, which treats augmented versions of an image as a positive pair and other images as negative samples. While most studies related to contrastive learning deal with unsupervised learning, Khosla et al. [35] extended contrastive learning to a fully supervised setting to leverage label information by considering all samples from the same class as positive.

There are several works where the contrastive learning strategy is applied to the semantic segmentation task. Chaitanya et al. [36] proposed contrastive strategies that leverage structural similarities across volumetric medical images under self-supervised learning. This approach treats representations of slices from similar volumes as positive pairs and representations of slices from dissimilar volumes as negative pairs. Zhao et al. [37] proposed a fine-tuning approach after pretraining the encoder using contrastive loss on semantic segmentation tasks. However, these works were motivated to achieve better performance with a small amount of training data rather than train a segmentation network encoder to learn the semantic embedding property.

Concurrently with our work, Wang et al. [38] proposed using contrastive loss to the embedded features to learn semantic relations between features across different images. Their core idea is similar to ours, but they contrast the transformed features by projection layers and employ different sampling strategies. In contrast, we apply contrastive loss to the encoder’s features directly and verify the additional benefit, robustness to domain shifts.

**III. METHODS**

**A. FRAMEWORK OVERVIEW**

The overall framework of our proposed method is illustrated in Fig. 2. In our work, the objective of semantic segmentation is to obtain a binary segmentation map \( \hat{y} \in \mathbb{R}^{H \times W} \) that consists of class predictions for each pixel of the input image \( x \in \mathbb{R}^{H \times W} \). To enable a model to learn the semantic embedding property, we introduce the supervised contrastive loss for feature vectors extracted from the encoder. Our proposed contrastive loss aligns positive feature vectors closely with each other and separately from the negative feature vectors in the embedding space. To this end, we use the resized ground-truth masks for supervision as depicted in Fig. 2. Each feature vector in the feature map represents a specific patch in the input image. Therefore, each pixel of the resized label, which has the same size as the feature map, indicates...
whether the patch represented by the corresponding feature vector belongs to a positive or negative class.

To prevent the loss of information caused by an overly down-sampled label, we first expand the size of the feature map using bilinear interpolation and then resize the ground-truth label to the same size as the interpolated feature map. It is worth noting that our proposed contrastive loss is computed on a mini-batch and not an individual image basis to encourage a model to learn the representations that are invariant across images. However, computing the contrastive loss might be computationally too expensive to consider every possible pair of feature vectors in a given mini-batch. Hence, we present the boundary-aware sampling approach that effectively samples valuable subsets of feature vectors in terms of semantic embedding.

**B. CONTRASTIVE LOSS**

Since semantic segmentation is inherently a pixel-level classification task, the commonly used loss function for binary segmentation is the binary cross entropy (BCE) loss. For a predicted segmentation map $\hat{y}$ with a total of $N$ pixels and the ground-truth label $y$, BCE loss is computed as:

$$L_{BCE}(\hat{y}, y) = -\frac{1}{N} \sum_{i} (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)),$$

where $y_i \in \{0, 1\}$ and $\hat{y}_i \in [0, 1]$ are the ground-truth class and the predicted probability for the $i$-th pixel, respectively. Also, one can consider the Dice loss as a segmentation loss, which can be computed as:

$$L_{dice}(\hat{y}, y) = 1 - \frac{2 \sum_{i} \hat{y}_i y_i}{\sum_{i} \hat{y}_i^2 + \sum_{i} y_i^2}.$$

In this work, we used a combination of BCE loss and Dice loss for segmentation loss [39].

We argue that these commonly used loss functions do not guarantee that the encoder of a segmentation network will learn the semantic embedding property as validated in Section IV-D. To obtain an encoder that has this desirable property, we apply the contrastive loss to the embedded feature vectors. For a feature map $Z \in \mathbb{R}^{C_f \times H_f \times W_f}$ extracted from an input image $x \in \mathbb{R}^{H \times W}$, the interpolated feature map is $Z' \in \mathbb{R}^{C_f \times H' \times W'}$ where $H' = cH_f$ and $W' = cW_f$ with an up-sampling factor $c$. Accordingly, the resized label will be $y' \in \mathbb{R}^{H' \times W'}$, and its $i$-th pixel is assigned as a label to the $i$-th vector $z_i \in \mathbb{R}^{C_f}$ in the interpolated feature map $Z'$. For a single positive feature vector $z_i$ in a mini-batch $B$, the proposed contrastive loss is defined as follows:

$$L_{con}(z_i, \{z^+\}_{-i}, \{z^-\}) = -\frac{1}{|\{z^+\}_{-i}|} \log \frac{\sum_{z_j \in \{z^+\}_{-i}} \exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{z_j \in \{z^+\}_{-i}} \cup \{z^-\} \exp(\text{sim}(z_i, z_j)/\tau)},$$

where $\{z^+\}$ and $\{z^-\}$ respectively are a set of positive and negative feature vectors in $B$. Here, $\text{sim}(\cdot, \cdot)$ denotes a cosine similarity function, which measures the similarity between two feature vectors and $\tau$ represents a temperature parameter. In this work, we set the temperature $\tau$ to 1 without tuning the hyperparameter.

It is important to note that $L_{con}$ is computed over a mini-batch rather individual images, i.e., $|\{z^\ast\} + \{z^-\}| = H' \times W' \times |\mathcal{B}|$. When $L_{con}$ is considered for each image, the model will only learn the relationship between positive and negative features in each image independently. To learn this relationship across all training samples, our proposed contrastive loss $L_{con}$ is computed over all images in a mini-batch. Hence, $L_{con}$ is minimized when a given positive feature is aligned with other positive features (i.e., the cosine similarities are nearly ones) in the representation space and orthogonal to all negative features (i.e., the cosine similarities are nearly zeros). It can be inferred from (3) that $L_{con}$ does not guarantee that the negative features will be located closely in the representation space. This is reasonable since the backgrounds in medical images (i.e., negative patches) usually contain heterogeneous characteristics while foregrounds such as lesions or organs (i.e., positive patches) are relatively homogeneous.

Consequently, we define the total loss function $L_{total}$ for a given mini-batch $B$ as follows:

$$L_{total} = \frac{1}{|\mathcal{B}|} \sum_{k} L_{BCE}(\hat{y}_k, y_k) + \frac{1}{|\mathcal{B}|} \sum_{k} L_{dice}(\hat{y}_k, y_k) + \frac{\lambda}{|\{z^+\}|} \sum_{z_i \in \{z^+\}} L_{con}(z_i, \{z^+\}_{-i}, \{z^-\}),$$

where $\hat{y}_k$ and $y_k$ are the prediction result and ground-truth label of $k$-th sample, and $\lambda$ is a weight for the contrastive loss and is set to 1 in all experiments. Training a segmentation model with $L_{total}$ encourages the encoder to have the semantic embedding property and the decoder to predict segmentation results based on the well-embedded features.

**C. BOUNDARY-AWARE SAMPLING**

Computing the proposed contrastive loss in (4) for all feature pairs in a mini-batch is computationally expensive since the number of pairs is too large even with a moderately sized mini-batch. For example, given the encoded feature map with a spatial resolution of $r \times r$ and the mini-batch size of $b$, the number of pairs to be examined is $\frac{1}{2}r^{2b}(r^{2b} - 1)$. Therefore, in practice, the feature vectors must be sampled to compute the contrastive loss.

One can simply adopt a random sampling strategy. However, a more effective sampling strategy can be designed based on the boundary regions generally being more challenging to segment correctly. In terms of learning the semantic embedding property of the encoder, it is helpful for it to focus on hard-to-classify regions rather than those easy-to-classify. Hence, we propose an effective sampling method,
boundary-aware sampling, which gives more attention to the feature vectors corresponding to the boundary regions.

Suppose that the number of positive $s_p$ and negative $s_n$ feature samples are given. Note that it is known that $s_p$ should be much larger than $s_p$ for contrastive learning [20] and this is also shown in our experimental results. The proposed boundary-aware sampling can be applied in three ways according to how $s^b_p$, the number of positive feature samples associated with the boundary area, is determined: fixed, random, and linear sampling.

- Fixed : Using a fixed proportion to decide the number of boundary features, i.e., $s^b_p = \alpha s_p$ where $\alpha \in [0, 1]$. In our experiments, we manually set $\alpha$ to 0.2.
- Random : Using a random proportion to decide the number of boundary features, i.e., $s^b_p = \alpha_s s_p$ where $\alpha \sim \text{Unif}(0, 1)$. Here, $\alpha$ is randomly sampled at each training iteration.
- Linear : Using a variable proportion that linearly increases according to the epoch, i.e., $s^b_p = \alpha s_p = \frac{t}{T}s_p$ where $t$ and $T$ represent the current and total number of epochs, respectively. We can expect the curriculum learning effect, which progressively focuses on the hard examples.

Note that the parameter $\alpha$ for random or linear sampling is determined automatically at each iteration whereas that for fixed sampling should be defined a priori. For all three different sampling strategies, the number of negative feature samples representing the boundary area $s^b_n$ is set to $s^b_p$. Consequently, the feature vectors corresponding to non-boundary areas are sampled as the amount of $s_p - s^b_p$ and $s_n - s^b_n$ for positive and negative samples, respectively. To demonstrate the efficacy of the proposed sampling methods, we examine performance gains according to the sampling methods in Section IV.

IV. EXPERIMENTS AND RESULTS

We performed two sets of experiments to demonstrate the efficacy of our proposed method regarding segmentation (referred to as source segmentation) and domain robustness performance. We show that applying contrastive loss with boundary-aware sampling yields better segmentation performance on liver and brain tumor segmentation tasks across the state-of-the-art architectures. Also, we provide quantitative and qualitative analyses to show that the encoder of the standard segmentation model does not learn the semantic embedding property. Additionally, for domain robustness, we present experimental results that our method also improves the segmentation model’s generalization performance on the unseen domain for lung and spinal cord gray matter segmentation tasks.

A. DATASETS

1) Liver Segmentation

Liver Tumor Segmentation Challenge (LiTS) dataset [40] consists of 201 3D abdominal CT scans with annotations on the liver and tumor regions. We used 131 CT scans that have publicly available labels and we focused on the liver segmentation task by only considering the liver as a positive class and the remainder as a negative class. We truncated the pixel intensity values of all scans to the range of [-200, 200] Hounsfield units (HU) to remove irrelevant details and enhance the contrast between liver and other tissues. Then, we normalized them to the value of [0, 255]. For training the segmentation models in our experiments, we used 2D slice images from the 3D CT scans. In total, we used 19,163 slices with $512 \times 512$ resolution except for the images that do not show liver. We randomly selected 103 CT scans as training data and the remaining 28 CT scans as test data.

2) Brain Tumor Segmentation

The BraTS 2018 training dataset [41] provides multimodal 3D brain MRIs annotated with three ground-truth segmentation labels: necrotic and non-enhanced tumor, edema, and enhanced tumor. The dataset contains 210 high-grade (HG) and 75 low-grade (LG) cases, and each case has four MRI modalities: FLAIR, T1, T1c, and T2. We randomly selected 20 cases for test data and the remaining cases were used as training data. For simplicity, we converted it into a binary problem by considering all three labels as a positive class and the remaining background as a negative class, following [9]. In the training process, we used 2D slice images from original 3D scans as inputs. In the evaluation process, we measured the segmentation performance on 3D scans by stacking 2D slice-level predictions.

3) Lung Segmentation

To evaluate the generalization performance on the unseen domain, we used three chest X-ray datasets acquired from different sources: Japanese Society of Radiological Technology (JSRT) [42], Montgomery Country (MC) [43], and Shenzhen (SZ) [43]. The JSRT dataset consists of 247 chest X-ray images, of which 154 cases have lung nodules and 93 cases are without lung nodules. MC dataset contains 138 X-rays from Montgomery Country’s tuberculosis screening program including 80 normal cases and 58 cases with manifestations of tuberculosis. The JSRT and MC datasets have manually segmented lung masks for evaluating automatic lung segmentation methods. However, the SZ dataset is not officially provided with ground-truth segmentation labels. Therefore, we referred to labels released by [44] and used 521 images by excluding the cases without annotations or that are mislabeled. Since the image acquisition equipment and lung area’s visual characteristics are different, we consider each dataset as an individual domain. Two datasets were used as source domains for training and the remaining dataset was used as the target domain (i.e., unseen domain). Specifically, we randomly selected 30% of each source domain to evaluate segmentation performance on the source domains, and the remaining 70% was used for training.
To demonstrate the efficacy of SCE, we set the model trained on the lung segmentation, we adopted one domain as the target dataset. Since the dataset was acquired at four different time points, the performance generally degrades. The factors for brightness and contrast adjustments were used for the remaining segmentation tasks.

| Dataset | Architecture | Method       | Precision (%) | Recall (%) | Dice (%) | ACD (l) | ASD (l) | NLL (l) |
|---------|--------------|--------------|---------------|------------|----------|---------|---------|---------|
| U-Net   | Baseline     | 89.37 ± 0.64  | 85.41 ± 0.67  | 85.43 ± 0.40 | 6.30 ± 0.26 | 6.96 ± 0.37 | 0.099 ± 0.005 |
| SCE     | 92.22 ± 0.36** | 86.85 ± 0.40** | 87.57 ± 0.37** | 5.20 ± 0.29** | 5.51 ± 0.31** | 0.063 ± 0.003** |
| SCE-fixed | 92.10 ± 0.27** | 87.38 ± 0.35** | 87.90 ± 0.32** | 5.02 ± 0.18** | 5.31 ± 0.18** | 0.062 ± 0.003** |
| SCE-random | 91.85 ± 1.73** | 87.67 ± 0.69** | 88.02 ± 1.22** | 5.07 ± 0.51** | 5.41 ± 0.63** | 0.064 ± 0.019** |
| SCE-linear | 92.56 ± 0.48** | 87.90 ± 0.17** | 88.36 ± 0.28** | 4.85 ± 0.33** | 5.15 ± 0.35** | 0.054 ± 0.003** |
| LiTS    | Baseline     | 89.38 ± 0.56  | 85.88 ± 0.43  | 86.11 ± 0.49  | 5.91 ± 0.33  | 6.44 ± 0.39  | 0.089 ± 0.004 |
| U-Net++ | Baseline     | 91.00 ± 0.21  | 85.81 ± 0.64  | 86.41 ± 0.37  | 5.60 ± 0.16  | 5.92 ± 0.15  | 0.065 ± 0.004 |
| DeepLabV3+ | Baseline   | 91.14 ± 0.18  | 82.75 ± 0.24  | 85.70 ± 0.24  | 2.34 ± 0.11  | 2.39 ± 0.14  | 0.040 ± 0.002 |
| U-Net   | Baseline     | 90.81 ± 0.24  | 83.37 ± 0.18  | 85.95 ± 0.17  | 2.33 ± 0.17  | 2.42 ± 0.23  | 0.041 ± 0.005 |
| BraTS   | Baseline     | 92.13 ± 0.34**| 82.64 ± 0.32**| 85.88 ± 0.17  | 1.92 ± 0.14**| 1.93 ± 0.13**| 0.035 ± 0.000**|
| U-Net++ | Baseline     | 92.02 ± 0.12**| 82.78 ± 0.20**| 86.01 ± 0.14  | 1.98 ± 0.13**| 1.99 ± 0.13**| 0.035 ± 0.000**|
| Spinal Cord Gray Matter Segmentation

We used the spinal cord gray matter segmentation challenge dataset [45], which contains MRI axial slices of healthy spinal cord. Since the dataset was acquired at four different medical centers with different MRI systems, we considered the data collected from each site as an individual domain. The four domains, named site1, site2, site3, and site4, have 30, 133, 177, and 134 MRI axial slices, respectively, after the slices having no annotations were excluded. Similar to the lung segmentation, we adopted one domain as the target unseen domain to evaluate the domain generalization performance. We sampled 30% of each source domain as test data and the remaining data as training data. Since the positive regions were too small compared to the input size, we applied center cropping to the images before training.

B. IMPLEMENTATION DETAILS

To demonstrate the efficacy of SCE, we set the model trained with the segmentation loss as a baseline (i.e., λ = 0 in Eq. (4)). For the segmentation architectures, we employed U-Net [28] and U-Net++ [9], which are popular in medical image analysis, and DeepLabV3+ [7], the latest version of the DeepLab series.

For the liver segmentation task, ResNet34 [46] was used as the segmentation model’s encoder and ResNet50 was used for the remaining segmentation tasks. All models were trained using the SGD optimizer with the momentum of 0.9 and weight decay of 1e-4 for 120 epochs. The initial learning rate was set to 0.01, decayed at the 80-th and 100-th epochs by a factor of 0.1. For data augmentation, we applied color jittering that randomly adjusts the brightness and contrast of an input. The factors for brightness and contrast adjustments are both randomly sampled from Unif(0.6, 1.4) at each iteration. The batch size is one of the important factors that significantly affect the performance of deep neural networks. When using a large batch size for a small-scale dataset (e.g., the lung segmentation and spinal cord segmentation datasets in our experiments), the performance generally degrades since the advantages of stochastic gradient descent cannot be fully utilized. Thus, we determined the batch size for each domain.

2 We used the open source library for image segmentation based on PyTorch: https://github.com/qubvel/segmentation_models.pytorch
dataset based on the validation performance: we set the batch size to 64 for both liver and brain tumor segmentation, 32 for lung segmentation, and 16 for spinal cord gray matter segmentation.

For all tasks, the number of positive features to compute the contrastive loss was set to $\times 2$ of the batch size, and the number of negative samples was set to $\times 6$ of the number of positive samples based on the empirical findings from the ablation study (see Fig. 5). For the proposed method, we simply set the value of $\lambda$ in (4) to 1.0 for all experiments and manually set the parameter $\alpha$ to 0.2 in the case of fixed boundary-aware sampling. It should be noted that we did not tune the hyperparameters, the number of positive and negative features, and the weight $\lambda$, for each experiment to demonstrate that the proposed method is insensitive to those hyperparameters.

C. EVALUATION METRICS

We used six different metrics to evaluate the performance of the segmentation models from various perspectives: precision, recall, Dice coefficient, average contour distance (ACD), average surface distance (ASD), and negative log likelihood (NLL). The ACD and ASD are distance metrics that measure how far the predicted positive region is from the actual positive region. Let $s_i$ be the $i$-th pixel on the boundaries of predicted segmentation mask $S$ and $g_j$ be the $j$-th pixel on the boundaries of ground-truth mask $G$. Considering $n_S$ and $n_G$ as the total number of pixels in $S$ and $G$, respectively, ACD and ASD are defined as:

$$ACD(S, G) = \frac{1}{2} \left( \frac{\sum d(s_i, G)}{n_S} + \frac{\sum d(g_j, S)}{n_G} \right), \quad (5)$$

$$ASD(S, G) = \frac{1}{n_S + n_G} \left( \sum_i d(s_i, G) + \sum_j d(g_j, S) \right), \quad (6)$$

where $d(s_i, G)$ is the Euclidean distance between $s_i$ and the closest pixel on $G$. Furthermore, we use the NLL metric, which can be interpreted as a cross entropy loss, to identify the impact of applying contrastive loss on the quality of the estimated probability.

D. SOURCE SEGMENTATION

The comparative results on the LiTS and BraTS datasets are reported in Table 1. Our proposed method is referred to as SCE, SCE+fixed, SCE+random, or SCE+linear according to the boundary-aware sampling methods. Note that we did not search for the best-performing sampling strategy on each dataset since we aimed to demonstrate that our proposed contrastive loss with boundary-aware sampling strategies are robust and widely applicable to various imaging modalities. We report the means and standard deviations over five runs, and we conducted the one-tailed $t$-test to verify rigorously the performance improvement’s statistical significance.

The results on LiTS dataset show that the proposed semantic contrastive embedding considerably improves the baseline regardless of the network architecture. For example, SCE improves the Dice coefficient from 85.43%, 86.11%, and 86.41% to 87.57%, 88.34%, and 87.58% respectively for U-Net, U-Net++ and DeepLabV3+. In addition, we observe that applying contrastive loss with boundary-aware sampling further enhanced the segmentation performance. For U-Net++ and DeepLabV3+, in the majority of metrics, random and linear boundary-aware sampling achieved better performance than simple SCE. In the case of U-Net, linear boundary-aware sampling shows superior performance than other sampling methods. Note that these performance improvements are statistically significant. For the BraTS dataset, we also observe similar results: the proposed semantic contrastive embedding improves the segmentation models and the boundary-aware sampling further enhances their performance. For example, all performance metrics on U-Net benefit from the proposed contrastive loss, especially when combined with the linear boundary-aware sampling strategy. For U-Net++ and DeepLabV3+, the baseline trained with only the segmentation loss can be boosted by using the semantic contrastive embedding while their recall, precision,
FIGURE 4. Analysis of embedding characteristics: (a) the Dice score and contrastive loss curves and (b) a t-SNE visualization of randomly sampled feature vectors from each model. Each color denotes the category of the feature vectors: positive and negative vectors corresponding to the boundary area (positive-bd and negative-bd, respectively) and the non-boundary area (positive-not-bd and negative-not-bd, respectively).

FIGURE 5. Ablation study on LiTS with U-Net to assess the effect of changing the number of positive and negative samples.

and NLL values are further improved via our boundary-aware sampling method.

The experimental results confirm that, for superior segmentation models, learning the semantic embedding property is required and all state-of-the-art architectures can benefit from learning that property. The linear boundary-aware sampling method was selected for the subsequent experiments since it showed consistent performance improvement across all datasets and architectures.

For qualitative comparison, we provide some of the segmentation results of U-Net on the LiTS and BraTS datasets in Fig. 3. From the results on the LiTS dataset, we observe that applying contrastive loss with boundary-aware sampling decreases false positive predictions and improves the prediction on a very small lung region compared to the baseline. For example, in the second row of Fig. 3(a), U-Net trained with our proposed method correctly classifies all areas of the right side of the image that are misclassified as positives by the baseline. The similar results can be observed in the results on the BraTS dataset. For instance, the third row of Fig. 3(b) shows the case where the positive region is very small. In this case, SCE localized the positive region and performed even better with boundary-aware sampling, whereas the baseline could not find any positive area at all.

We conducted further analysis to identify whether the performance improvements are indeed due to the encoder’s embedding property. Fig. 4(a) shows the U-Net Dice score curves and contrastive loss on the LiTS training dataset. The baseline’s contrastive loss does not decrease as training progresses, whereas the Dice score increases. This implies that the general training scheme does not guarantee that encoders will acquire the semantic embedding property. Without any constraints on feature representation, the segmentation models learn how to produce precise segmentation results without attention to learning semantically meaningful representations.

To provide a more intuitive understanding of the embedding property, we visualize randomly sampled feature vectors from the encoder of baseline and SCE in Fig. 4(b) by using t-SNE. The baseline visualization shows that the positive and negative vectors are confused and the feature vectors representing the boundary area are barely distinguishable. However, the positive and negative feature vectors are significantly more separable when applying contrastive loss. When applying contrastive loss with boundary-aware sampling, the feature vectors representing boundary areas are well grouped according to their corresponding class. These results confirm...
that, as we intended, the proposed contrastive loss helps the encoder to learn the semantic embedding property, thereby inducing the reported improvements.

Two hyperparameters, the number of positive and negative feature samples, must be determined for SCE. We conducted an ablation study to investigate the impact of these hyperparameters on the segmentation performance. Figure 5 shows the precision, recall, and Dice scores of the proposed method with respect to the number of positive and negative samples from each mini-batch. The number of positive samples does not significantly affect the segmentation performance as shown in Fig. 5(a). Conversely, in Fig. 5(b), we see that the number of negative samples is important for the segmentation performance of SCE; when more negative samples are used, the performance improves. Note that this finding is consistent with the previous studies [20], [34] showed that a large number of negative samples is required for self-supervised contrastive learning.

E. DOMAIN ROBUSTNESS

We conducted additional experiments to verify that the proposed method can also be effective for improving domain robustness. Since SCE contrasts the feature vectors from different classes across inputs, it can be expected to learn domain invariant representations when trained with multi-source domains. Intuitively, this representation should improve the domain robustness of a model and may also improve performance on the unseen domain.

To examine the effect of the proposed method on domain robustness, we use lung segmentation and spinal cord segmentation datasets that consist of three and four independent domains, respectively. Specifically, one domain was adopted as the target domain and the remaining were considered as source domains. All models were trained on the source domains and evaluated on both the source and target domains. Note that the target domain is not accessible during training, which is then interpreted as an entirely unseen domain. We evaluated the baseline, SCE, and SCE with linear boundary-aware sampling by considering three representative measures including the overlap map Dice score, distance measure ACD, and probabilistic measure NLL.

1) Lung Segmentation

The comparison results on the lung segmentation task are summarized in Table 2. From the results, we observe that applying contrastive loss improves all performance measures on the target unseen domain. For example, in the setting of “JSRT, MC (source domains) → SZ (target domain)”, the proposed method improves the Dice scores of U-Net, U-Net++, DeepLabV3+, from 94.37%, 94.51%, and 94.26% to 94.97%, 94.83%, and 94.58%, respectively. Also, the performances of U-Net++ and DeepLabV3+ were further enhanced by using linear boundary-aware sampling. In the setting of “MC, SZ → JSRT”, all architectures benefit from SCE especially DeepLabV3+. The results from the “SZ, JSRT → MC” setting lead to a similar conclusion, except that the effect of boundary-aware sampling is not clearly observed.

For source domains, the proposed method enhances the segmentation performance only in a few cases. We believe that it is because lung segmentation on CXRs is a relatively easy task and even the baseline produces a fairly high segmentation performance. Nevertheless, the proposed SCE strongly effects the robustness of models regarding domain generalization because the model can learn domain invariant representations.

The qualitative results of lung segmentation on the target domains are presented in Fig. 6(a). From the segmentation results on the target domains, compared to the baseline, we observe that applying contrastive loss reduces false positive regions that can be reduced even more with boundary-aware sampling. These results confirm that the proposed method can also improve segmentation performance on the unseen domain by contrasting positive and negative regions across multi-source domains.

2) Spinal Cord Gray Matter Segmentation

We next consider the spinal cord gray matter segmentation task to demonstrate the effectiveness of our proposed method on domain generalization. Table 3 summarizes the comparison results. From the results, we can see that the proposed method enhances segmentation performance in the majority
TABLE 3. Domain robustness for spinal cord gray matter segmentation

| Architecture | Method            | Source   | Target | Dice (%) | ACD | NLI |
|--------------|-------------------|----------|--------|----------|-----|-----|
|              |                   | site 2, 3, and 4 → site 1 |        |          |     |     |
| U-Net        | Baseline          | 88.07    | 0.91   | 0.016    | 84.75 | 1.16 | 0.021 |
|              | SCE               | 88.27    | 0.90   | 0.015    | 84.96 | 1.14 | 0.020* |
|              | U-Net++           | 88.84    | 0.90   | 0.015**  | 85.12 | 1.11*| 0.019* |
|              | Baseline          | 88.33    | 0.93   | 0.015    | 84.21 | 1.18 | 0.025 |
|              | SCE               | 88.37*   | 0.88** | 0.015**  | 85.39 | 1.07**| 0.019**|
|              | U-Net++           | 88.32**  | 0.88** | 0.016**  | 84.58 | 1.16**| 0.022**|
| DeepLabV3+   | Baseline          | 86.98    | 1.03   | 0.014    | 83.40 | 1.25 | 0.023 |
|              | SCE               | 87.20    | 1.00** | 0.017**  | 83.59 | 1.22**| 0.023** |
|              | U-Net++           | 87.11    | 1.01** | 0.017**  | 83.65 | 1.21**| 0.023** |
|              |                   | site 1, 3, and 4 → site 2 |        |          |     |     |
| U-Net        | Baseline          | 88.06    | 0.90   | 0.015    | 85.62 | 1.16 | 0.027 |
|              | SCE               | 88.37**  | 0.88** | 0.013**  | 85.37 | 1.14**| 0.024** |
|              | U-Net++           | 88.36**  | 0.87** | 0.014**  | 85.37 | 1.14**| 0.025** |
| DeepLabV3+   | Baseline          | 88.04    | 0.99   | 0.017    | 85.08 | 1.17 | 0.029 |
|              | SCE               | 88.35**  | 0.88** | 0.016    | 85.10 | 1.11**| 0.030** |
|              | U-Net++           | 88.42**  | 0.86** | 0.015**  | 85.18 | 1.13**| 0.031** |
| DeepLabV3+   | Baseline          | 87.69    | 0.99   | 0.017    | 83.97 | 1.27 | 0.028 |
|              | SCE               | 87.46**  | 0.96** | 0.015**  | 84.35 | 1.23**| 0.026** |
|              | U-Net++           | 87.37**  | 0.96** | 0.016**  | 84.29 | 1.23**| 0.026** |
|              |                   | site 1, 2, and 4 → site 3 |        |          |     |     |
| U-Net        | Baseline          | 88.25    | 0.99   | 0.016    | 86.60 | 2.03 | 0.022 |
|              | SCE               | 88.39    | 0.96** | 0.014**  | 85.06 | 2.07 | 0.020 |
|              | U-Net++           | 88.25    | 0.96** | 0.014**  | 79.86 | 2.41 | 0.021 |
| DeepLabV3+   | Baseline          | 88.28    | 0.94   | 0.017    | 78.28 | 2.63 | 0.028 |
|              | SCE               | 88.48    | 0.94   | 0.015**  | 80.79 | 1.93 | 0.030** |
|              | U-Net++           | 88.45    | 0.94   | 0.015**  | 79.26 | 2.96 | 0.033** |
| DeepLabV3+   | Baseline          | 86.81    | 1.11   | 0.019    | 80.25 | 1.79 | 0.022 |
|              | SCE               | 87.08**  | 1.09** | 0.018**  | 79.14 | 2.08**| 0.024** |
|              | U-Net++           | 87.07**  | 1.10** | 0.018**  | 79.41**| 1.91**| 0.024** |
|              |                   | site 1, 2, and 3 → site 4 |        |          |     |     |
| U-Net        | Baseline          | 87.22    | 0.92   | 0.017    | 86.68 | 1.08 | 0.024 |
|              | SCE               | 87.30    | 0.92   | 0.015**  | 86.10 | 1.05**| 0.021**|
|              | U-Net++           | 87.55**  | 0.90   | 0.015**  | 86.03 | 1.06**| 0.022**|
| DeepLabV3+   | Baseline          | 87.24    | 0.91   | 0.018    | 86.11 | 1.10 | 0.027 |
|              | SCE               | 87.58**  | 0.89** | 0.016    | 86.42 | 1.07**| 0.025 |
|              | U-Net++           | 87.76**  | 0.86** | 0.016**  | 86.24 | 1.08**| 0.025 |
| DeepLabV3+   | Baseline          | 85.65**  | 1.10** | 0.018**  | 88.04 | 1.02**| 0.017**|
|              | SCE               | 85.53**  | 1.10** | 0.018**  | 87.97 | 1.03**| 0.017**|

Bold-faced: where the proposed SCE shows statistically better results than the baseline. * p-value < 0.1; ** p-value < 0.05 compared with the baseline. ∗ p-value < 0.01; ∗∗ p-value < 0.05 compared with the simple SCE.

V. CONCLUSION

In this work, we argue that typical training schemes for deep segmentation networks do not ensure the desirable semantic embedding property of the encoder-decoder networks. To enable the encoder to learn this property during training, we introduce a novel framework that utilizes supervised contrastive learning for the segmentation task. We extend the existing contrastive loss to contrast feature representations from the encoder by assigning the feature-level labels derived from the ground-truth masks. In addition, we propose a boundary-aware sampling strategy that samples feature vectors in a given mini-batch that corresponds to the boundary area, which can be treated as a particularly informative area for improving segmentation performance. The extensive experiments on liver and brain tumor segmentation with popular architectures show that across several metrics, the proposed SCE method effectively improves segmentation performance. Through further analyses of training curves and feature visualization, the semantic embedding property is plugging. These results, which are consistent with other qualitative results, confirm the advantages of our proposed method.
confirmed to have produced the performance improvement. We conducted further experiments on lung and spinal cord gray matter segmentation to demonstrate that the proposed method boosts the domain robustness of segmentation networks by learning domain invariant representations. From the results, we confirm that segmentation models trained with the proposed method are more robust to domain shift.

Although the proposed method clearly shows the advantages in terms of segmentation and domain robustness, there exist some limitations. The first is that the computational cost is slightly increased. We tried to mitigate this issue through boundary-aware sampling, but it still needs additional computations than standard training schemes. Another limitation is that boundary-aware sampling may not be effective for some datasets with small regions of interest when a segmentation network has a large receptive field size. These limitations are valuable to be addressed in future research.

REFERENCES

[1] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” nature, vol. 542, no. 7639, pp. 115–118, 2017.

[2] J. G. Nam, S. Park, E. J. Hwang, J. H. Lee, K.-N. Jin, K. Y. Lim, T. H. Yu, J. H. Sohn, S. Hwang, J. M. Goo, and C. M. Park, “Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs,” Radiology, vol. 290, no. 1, pp. 218–228, 2019.

[3] H.-E. Kim, H. H. Kim, B.-K. Han, K. H. Kim, K. Han, H. Nam, E. H. Lee, and E.-K. Kim, “Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multicenter study,” The Lancet Digital Health, vol. 2, no. 3, pp. e138–e148, 2020.

[4] T. Zhou, S. Ruan, and S. Canu, “A review: Deep learning for medical image segmentation using multi-modality fusion,” Array, vol. 3, p. 100004, 2019.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pp. 834–848, 2017.

[6] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” arXiv preprint arXiv:1706.05587, 2017.

[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.

[8] Z. Zhang, C. Wu, S. Coleman, and D. Kerr, “Dense-inception u-net for medical image segmentation,” Computer methods and programs in biomedicine, vol. 192, p. 105395, 2020.

[9] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: Redesigning skip connections to exploit multiscale features in image segmentation,” IEEE Transactions on Medical Imaging, vol. 39, no. 6, pp. 1856–1867, 2020.

[10] N. Itzhaz and M. S. Rahman, “Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation,” Neural Networks, vol. 121, pp. 214–227, 2020.

[11] J. Chen, Y. Li, Q. Xu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “Transunet: Transformers make strong encoders for medical image segmentation,” arXiv preprint arXiv:2102.04306, 2021.

[12] Z. Hayden, X. He, and M. Salzmann, “Boundary-aware instance segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6506–6507.

[13] K. C. Wong, M. Moradi, H. Tang, and T. Syeda-Mahmood, “3d segmentation with exponential logarithmic loss for highly unbalanced object sizes,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2018, pp. 612–619.

[14] N. Abraham and N. M. Khan, “A novel focal tversky loss function with improved attention u-net for lesion segmentation,” in 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, 2019, pp. 683–687.

[15] W. Zhu, Y. Huang, L. Zeng, X. Chen, Y. Liu, Z. Qian, N. Du, W. Fan, and X. Ye, “Anatomically-aware deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy,” Medical physics, vol. 46, no. 2, pp. 576–589, 2019.

[16] H. Kervadec, J. Bouchitba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, “Boundary loss for highly unbalanced segmentation,” in International conference on medical imaging with deep learning. PMLR, 2019, pp. 285–296.

[17] C. Paetzoldt, S. Shi, I. Ezhov, G. Tetteh, A. Erturk, H. Z. Munich, and B. Menze, “cldice—a novel connectivity-preserving loss function for vessel segmentation,” in Medical Imaging Meets NeurIPS 2019 Workshop, 2019.

[18] S. Shi, J. C. Paetzoldt, A. Sekuboyina, A. Zhylka, I. Ezhov, A. Unger, J. P. Pluim, G. Tetteh, and B. H. Menze, “cldice—a topology-preserving loss function for tubular structure segmentation,” arXiv preprint arXiv:2003.07311, 2020.

[19] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06), vol. 2. IEEE, 2006, pp. 1735–1742.

[20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in International conference on machine learning. PMLR, 2020, pp. 1597–1607.

[21] M. Wang and W. Deng, “Deep visual domain alignment: A survey,” Neurocomputing, vol. 312, pp. 135–153, 2018.

[22] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, B. J. Wood, H. Roth, A. Myronenko, D. Xu et al., “Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation,” IEEE transactions on medical imaging, vol. 39, no. 7, pp. 2531–2540, 2020.

[23] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, H. Roth, A. Myronenko, D. Xu, and Z. Xu, “When unseen domain generalization is unnecessary? rethinking data augmentation,” arXiv preprint arXiv:1906.03347, 2019.

[24] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[25] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” arXiv preprint arXiv:1511.00561, 2016.

[26] H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1520–1528.

[27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected crfs,” arXiv preprint arXiv:1412.7062, 2014.

[28] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[29] L. Lin, J. Cheng, Q. Quan, F.-X. Wu, Y.-P. Wang, and J. Wang, “A survey on u-shaped networks in medical image segmentations,” Neurocomputing, vol. 409, pp. 244–258, 2020.

[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollàr, “Focal loss for dense object detection,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[31] C. H. Sudre, L. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, “Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations,” in Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, 2017, pp. 240–248.

[32] S. M. Salehi, D. Erdogmus, and A. Gholipour, “Tversky loss function for image segmentation using 3d fully convolutional deep networks,” in International workshop on machine learning in medical imaging. Springer, 2017, pp. 379–387.

[33] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’20), 2020, pp. 9729–9738.

[34] X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” arXiv preprint arXiv:2003.04297, 2020.
[35] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, “Supervised contrastive learning,” arXiv preprint arXiv:2004.11362, 2020.

[36] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, “Contrastive learning of global and local features for medical image segmentation with limited annotations,” arXiv preprint arXiv:2006.10511, 2020.

[37] X. Zhao, R. Vemulapalli, P. Mansfield, B. Gong, B. Green, L. Shapira, and Y. Wu, “Contrastive learning for label-efficient semantic segmentation,” arXiv preprint arXiv:2012.06985, 2020.

[38] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. V. Gool, “Exploring cross-image pixel contrast for semantic segmentation,” arXiv preprint arXiv:2101.11939, 2021.

[39] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wassermann, G. Koehler, T. Narajjitra, S. Wirkert et al., “nnu-net: Self-adapting framework for u-net-based medical image segmentation,” arXiv preprint arXiv:1809.10486, 2018.

[40] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser et al., “The liver tumor segmentation benchmark (lls),” arXiv preprint arXiv:1901.04056, 2019.

[41] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki et al., “Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge,” arXiv preprint arXiv:1811.02629, 2018.

[42] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujiyama, Y. Kodera, and K. Doi, “Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules,” American Journal of Roentgenology, vol. 174, no. 1, pp. 71–74, 2000.

[43] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wang, P.-X. Lu, and G. Thoma, “Two public chest x-ray datasets for computer-aided screening of pulmonary diseases,” Quantitative imaging in medicine and surgery, vol. 4, no. 6, p. 475, 2014.

[44] S. Sirenko, Y. Kochura, O. Alienin, O. Rokovyi, Y. Gordienko, P. Gang, and W. Zeng, “Chest x-ray analysis of tuberculosis by deep learning with segmentation and augmentation,” in 2018 IEEE 38th International Conference on Electronics and Nanotechnology (ELNANO). IEEE, 2018, pp. 422–428.

[45] F. Prados, J. Ashburner, C. Blaiotta, T. Brosch, J. Carballido-Gamio, M. J. Cardoso, B. N. Conrad, E. Datta, G. Dávid, B. De Leener et al., “Spinal cord grey matter segmentation challenge,” Neuroimage, vol. 152, pp. 312–329, 2017.

[46] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

SANGWOO LEE received the B.S. degree in the Department of Industrial and Information Systems Engineering from Seoul National University of Science and Technology, Seoul, South Korea in 2021. He is currently pursuing the M.S. degree in the Department of Data Science, Seoul National University of Science and Technology, Seoul, South Korea. His research interests include out-of-domain robustness and self-supervised representation learning with deep neural networks.

YEJIN LEE received the B.S. degree in the Department of Industrial and Information Systems Engineering from Seoul National University of Science and Technology, Seoul, South Korea in 2020. She is currently pursuing the M.S. degree in the Department of Data Science, Seoul National University of Science and Technology, Seoul, South Korea. Her research interests include deep learning, image segmentation, and out-of-domain robustness.

GEONGYU LEE received the B.S. degree in the Department of Information Security from Daejeon University in 2019 and the M.S. degree in the Department of Data Science from the Seoul National University of Science and Technology in 2021. He is currently working at DeepBio Inc.

SANGHEUM HWANG received B.S. and Ph.D. degrees in industrial engineering from KAIST, Daejeon, Korea, in 2005 and 2012, respectively. Currently, he is an Assistant Professor in the Department of Industrial and Information Systems Engineering and the Research Center for Electrical and Information Technology, Seoul National University of Science and Technology, Seoul, Korea. His research interests are in the areas of statistical learning algorithms, and deep learning and its practical applications.